

**TENNESSEE**

**July 16, 2001 Filing**

**LIST OF EXHIBITS TO  
BELLSOUTH DIRECT TESTIMONY  
IN  
DOCKET 01-00193**

**Mulrow**

EJM-1

EJM-2

**Coon**

DAC-1

DAC-2

DAC-3

DAC-4

DAC-5

1                               BELLSOUTH TELECOMMUNICATIONS, INC.  
2                               DIRECT TESTIMONY OF EDWARD J. MULROW, PH.D.  
3                               BEFORE THE TENNESSEE REGULATORY AUTHORITY  
4                               DOCKET NO. 01-00193  
5                               JULY 16, 2001  
6

7    Q.     PLEASE STATE YOUR NAME, WHO YOU WORK FOR, AND YOUR  
8            BUSINESS ADDRESS.  
9

10   A.    My name is Edward J. Mulrow. I am employed by Ernst & Young LLP as a  
11           Senior Manager in the Quantitative Economics and Statistics Group. I have been  
12           retained by BellSouth as a statistical advisor. My business address is 1225  
13           Connecticut Ave., NW, Washington, DC 20036.  
14

15   Q.     WHAT IS YOUR PROFESSIONAL EXPERIENCE AND EDUCATIONAL  
16            BACKGROUND?  
17

18   A.    My career as a statistical consultant spans over 13 years. While at Ernst &  
19           Young, I have been involved in a number of regulatory issues for several  
20           telecommunications companies. Prior to my employment at Ernst & Young, I was  
21           a senior scientist at Science Applications International Corporation (SAIC) where  
22           I was involved in the analyses of current and future defense systems. I also have  
23           worked as a senior sampling statistician at the National Opinion Research Center  
24           (NORC) at the University of Chicago, a mathematical statistician for the Internal  
25           Revenue Service, and an assistant professor of mathematics for Southern Illinois

1 University. I received a BA in mathematics from Illinois Wesleyan University, an  
2 MS in mathematics from the University of Utah, and a Ph.D. in statistics from  
3 Colorado State University.

4  
5 Q. WHAT IS THE PURPOSE OF YOUR TESTIMONY?

6  
7 A. I am here to address statistical issues for this docket. I will speak to issues  
8 involving the appropriate methodology for determining whether BellSouth is  
9 providing parity: 1) to individual CLECs (Tier I), and 2) to the CLEC community  
10 as a whole (Tier II). I will also discuss issues related to BellSouth's penalty  
11 payment calculation.

12  
13 Q. PLEASE SUMMARIZE YOUR TESTIMONY.

14  
15 A. I generally agree with the statistical methodology that the Tennessee Regulatory  
16 Authority (the Authority) ordered to be used in the ITC^Deltacom arbitration.  
17 The key points with which I agree are:

- 18  
19 1. The Truncated Z statistical test should be used when transaction level data is  
20 available and a BellSouth retail analog exists.  
21 2. The statistical testing methodology should balance Type I and Type II error  
22 probabilities.  
23 3. The same methodology should be used for both Tier I and Tier II testing.  
24

1 I will address each of these points in more detail in my testimony. The  
2 methodology that I am recommending was developed in a joint effort with AT&T  
3 statistician Dr. Colin Mallows (now retired). While the joint CLEC coalition has  
4 not adopted the methodology at this point in time, I believe that the records of the  
5 recent Florida and North Carolina performance measure hearings reflect that  
6 CLEC coalition expert witness, AT&T's Dr. Robert Bell, and I agree that the  
7 statistical tests offered by each party are correct, and that the choice of the  
8 statistical test that will be used, will be determined by the elements of the penalty  
9 plan that the Authority selects. If the Authority selects a plan that uses an  
10 aggregate statistical test, as it did in the Deltacom arbitration, then the Truncated  
11 Z methodology proposed by BellSouth is appropriate.

12  
13 Q. CAN YOU PROVIDE A BRIEF OVERVIEW OF WHAT WE ARE TRYING  
14 TO ACCOMPLISH WITH THE STATISTICAL ANALYSIS THAT YOU ARE  
15 GOING TO DESCRIBE IN YOUR TESTIMONY?

16  
17 A. Yes. What we are talking about here is the situation where BellSouth provides a  
18 service of some sort to its competitors, the CLECs. BellSouth also, at the same  
19 time, is providing a similar, or at least an analogous service, to its own retail  
20 operations. The question is whether BellSouth is favoring its retail operations in  
21 the provision of the particular service, or whether it is providing the same level of  
22 service to its competitors as it provides to itself.

23  
24 For instance, assume that CLECs purchased widgets from BellSouth and  
25 BellSouth also provided widgets to its own retail operations which then used the

1 widgets to provide service to BellSouth's own retail customers. If BellSouth  
2 provided the widgets to the CLECs on a two-day interval every time, and provided  
3 the widgets to its own retail operations on a two-day interval every time, then  
4 anyone could conclude that BellSouth was providing parity to the CLECs.

5

6 Similarly, if BellSouth were furnishing the widgets to the CLECs on a one-day  
7 interval, and furnishing the widgets to its own retail operations in two days, it  
8 would be evident that BellSouth wasn't providing parity, but was providing better  
9 service to the CLECs than to its own retail operations. Presumably the CLECs  
10 would not be upset with that.

11

12 The problem arises when BellSouth, in a given month, provides the widgets to its  
13 retail operations on average in two days, and provides widgets to the CLECs, on  
14 average, in 2.2 days. The question is whether the difference is attributable to  
15 random chance, or whether the difference is attributable to either some systemic  
16 problem with BellSouth's operations or some intentional act on BellSouth's part.  
17 The purpose of the statistical analysis is to provide the tools that the Authority can  
18 use to make an informed judgment about whether the difference I just described is  
19 something to be concerned about or rather is simply the result of the sample used  
20 and therefore meaningless. The specific tool that I am going to describe in my  
21 testimony is a test that can be applied whenever the Authority wishes to compare  
22 two outcomes to determine whether any perceived difference in the outcomes is  
23 real or not. While the test is a statistical one, and involves statistical concepts, I  
24 believe that what we have is very workable and understandable.

25

1 Q. WHAT IS THE APPROPRIATE STATISTICAL METHODOLOGY THAT  
2 SHOULD BE EMPLOYED TO DETERMINE IF BELL SOUTH IS PROVIDING  
3 COMPLIANT PERFORMANCE?  
4

5 A. The appropriate methodology to use is called the Truncated Z method with error  
6 probability balancing. Dr. Colin Mallows, a recently retired statistician from  
7 AT&T Research Labs, created the Truncated Z statistic, and then Dr. Mallows  
8 together with Ernst & Young statisticians, including myself, developed the actual  
9 Truncated Z methodology. The methodology is distinguished from the statistic in  
10 that we jointly took Dr. Mallows' formula that yielded the statistic and  
11 complemented it with such things as the error probability balancing. The  
12 collaborative effort was the result of a request by the Louisiana Public Service  
13 Commission (LPSC), lasted over nine months, and concluded in the filing of a  
14 "Statisticians' Report" with the LPSC in September of 1999 (revised February  
15 2000 -- attached as Exhibit No. EJM-1).<sup>1</sup>  
16

17 Q. CAN YOU EXPLAIN IN LAYMAN'S TERMS WHAT THE TRUNCATED Z  
18 METHODOLOGY DOES?  
19

20 A. I can. Remember that what we are doing is comparing two outcomes to see if  
21 there is any difference. Therefore, one of the first things that must be done is to  
22 separate all of our observations into identical, or substantially identical categories.  
23 For instance, let's assume that what we are trying to compare is the performance  
24 of BellSouth with regard to order completion intervals. That is, we want to know  
25 whether the order completion intervals for BellSouth's retail operations are

---

<sup>1</sup> Typographical error corrections are attached as Exhibit No. EJM-2.

1 statistically the same as the order completion intervals for the CLECs. You would  
2 not want to compare a BellSouth retail residential order that requires a dispatch  
3 with a CLEC resale residential order that did not require a dispatch. The  
4 requirements for provisioning the different orders would be different.

5  
6 Obviously, you can carry this concept of granularity to an extreme, but the point is  
7 that the first thing we have to do is to separate the individual observations into  
8 enough categories so that the comparison we are going to make is as close to  
9 being an apples-to-apples comparison as we can reasonably get it.

10

11 In our work, we call these classifications “cells.” For any particular measurement  
12 contained in the BellSouth plan, there could be thousands of these “cells.” Once  
13 we have these cells identified and populated with observations, we apply  
14 statistical tests to the information in the cells to put the conclusions we draw about  
15 every cell on a common footing. To make this illustration as clear as possible, I  
16 will assume that I have a cell for residential dispatched orders during the first half  
17 of the month. For illustrative purposes, I will assume that BellSouth has one  
18 observation that took 2 days, and the CLECs had a single observation that took  
19 2.2 days, the times I used above. We would then apply a statistical calculation to  
20 those two observations, as is described in Appendix A of Exhibit EJM-1, and we  
21 would derive a “cell z-value” of -0.67. The calculation of this value is not subject  
22 to a simple explanation, but is done through standard statistical analysis with  
23 which no statistician should disagree. Obviously, as the number of observations  
24 in the cell increases, the “cell z-value” may change.

25

1 I have described briefly what we would do for the individual cell. In actuality, we  
2 would make this same type of calculation for every cell (or more plainly stated,  
3 for each of the apples-to-apples comparisons that we had identified in connection  
4 with the specific measurement).

5

6 Q. WHAT HAPPENS NEXT?

7

8 A. When we are done, we would have a large number, potentially thousands of  
9 numbers, each representing the “cell z-value” for each individual cell. The “cell  
10 z-values” would be either positive, or negative, or in some cases would be zero.  
11 The cells that have a negative “cell z-value” would represent those cells where,  
12 continuing my example from above, it appears that the interval for the CLECs was  
13 longer than for BellSouth. The cells that had a positive “cell z-value” would  
14 represent those cells where, again continuing my example, it appears that the  
15 interval for the CLECs was shorter than for BellSouth. Where the “cell z-value”  
16 was zero, there would be no apparent difference in the intervals.

17

18 Q. WHAT DO YOU DO WITH THESE THOUSANDS OF “CELL Z-VALUES?”

19

20 A. We move to the next step in the analysis, which is to analyze the “cell z-values”  
21 using a normal distribution curve. If BellSouth were providing parity, one would  
22 expect that the distribution of the values over the entire range of the cells would  
23 look just like the normal bell curve with which we should all be familiar.

24



1 This is where the idea of “truncating” the z statistic comes into play. We have z  
2 statistics for every cell. Some are positive, meaning they fall on the right side of  
3 the normal bell curve. Some are negative, which means that they are on the left  
4 side of the normal bell curve. One concern we would have is that if all of the z-  
5 values were left in the analysis, the positive z-values, if there were enough of  
6 them, might mask one or more significant negative z-values when averaging the z-  
7 values across all cells. That is, if there were a thousand cells, and 800 of them had  
8 positive z statistics, the sheer number of positive observations might hide  
9 significant negative values. Therefore, in order to prevent this, the Truncated Z  
10 methodology simply sets every positive value to zero, hence the “truncation.” By  
11 setting the positive observation to zero, it forces us to concentrate on the negative  
12 values on the left side of the bell shaped curve.

13  
14 Q. WHAT DO YOU DO NEXT?

15  
16 A. Remember we are now only concentrating on the lower half of the normal bell-  
17 shaped curve, and what we are going to try to do, in layperson’s terms, is to  
18 determine how far the observations we have made fall from the normal bell curve  
19 I have been talking about. You would not expect the observations to lie down  
20 perfectly on the curve. There are going to be variations and the question is how  
21 much is too much. Consequently, the next step is to calculate a Z statistic for all  
22 the cells, including those formerly positive cells whose value has now been set to  
23 zero. Assuming that a statistician understood the purpose of truncating the  
24 positive values, and the selection of the cells weights, the calculation of the Z  
25 statistic for the truncated observations (the positive ones set to zero and the

1 remaining negative observations left as they were found) should not be subject to  
2 dispute. This calculation will leave you with a single number that represents the  
3 Truncated Z statistic value for the particular measurement contained in  
4 BellSouth's plan for which the observations were made.

5

6 Q. DOES THIS CALCULATED Z STATISTIC BY ITSELF REPRESENT A  
7 STATISTICALLY SIGNIFICANT DIFFERENCE IN THE PERFORMANCE  
8 BELLSOUTH PROVIDED TO ITS RETAIL OPERATIONS AND THE CLECS?

9

10 A. No, generally you can't draw any conclusion from the Z statistic itself. It is just a  
11 number. However, if the number turns out to be positive (which, even though it  
12 seems illogical because of changing the positive values to zero, could occur) you  
13 could just ignore the result. If it is negative, however, you still have to have a  
14 number to compare the Z statistic to, in order to determine whether the difference  
15 represented by the Z statistic is significant.

16

17 Q. ONCE YOU HAVE THIS NEGATIVE Z STATISTIC, THEN, WHERE DO  
18 YOU GET THE NUMBER THAT IT IS COMPARED WITH IN ORDER TO  
19 DETERMINE WHETHER THERE IS A STATISTICALLY SIGNIFICANT  
20 DIFFERENCE IN THE SERVICE PROVIDED TO THE CLECS AND THE  
21 SERVICE BELLSOUTH PROVIDES TO ITSELF WITH REGARD TO THE  
22 SPECIFIC ITEM THAT YOU ARE MEASURING?

23

24 A. There are several ways of determining the number that is used for comparison.  
25 Given the constraints of a self-effectuating system, the best way, in my opinion, is

1 to use what we call “Error Probability Balancing.” Using this approach allows the  
2 observer to determine both that the observed difference is statistically significant,  
3 and that it is material. I will discuss this in more detail subsequently in my  
4 testimony.

5

6 Q. WHAT ARE SOME OF THE OTHER WAYS?

7

8 A. The most common statistical method used is what we call the “fixed critical  
9 value.” Let me explain what this is, and why it shouldn’t be used here. One of  
10 the main issues statisticians have to face in determining whether there is a  
11 statistical difference between two numbers is controlling the probability that the  
12 observed difference indicates a failure to provide parity when in fact parity has  
13 been achieved. We call these kind of errors, where it appears that there is a  
14 statistically significantly difference when there is in fact not one, a Type I error.  
15 To illustrate this point, consider the situation where a person is flipping a coin.  
16 Everyone knows that on average, heads should come up the same number of times  
17 as tails. Suppose you flip the coin five times, and just as a matter of chance, tails  
18 comes up every time. You might then conclude that something is wrong with the  
19 coin, that the coin is somehow biased toward tails because it is not acting in  
20 accord with what we know to be correct. In fact, the coin may be perfectly okay,  
21 and what we are seeing is simply a Type I error.

22

23 One way, then, to determine the “critical value” that is to be compared to the Z  
24 statistic that we have been talking about is to determine what the acceptable level  
25 of a Type I error is, and when that is done, a “critical value” can be calculated

1 using standard statistical tools. For instance, if you wanted the probability of a  
2 Type I error occurring limited to less than a 5 percent chance, the calculated  
3 “critical value,” based on a standard normal distribution, would be  $-1.645$ . Every  
4 statistician in the world would agree with the calculation of that number given the  
5 criteria we have laid out.

6

7 Q. WHAT WOULD YOU DO WITH THIS “CRITICAL VALUE” IF THAT WERE  
8 THE APPROACH TAKEN?

9

10 A. This is what is called a “fixed critical value.” All you would have to do is  
11 compare the Truncated Z statistic that we obtained as described above, with this  
12 value. If the Truncated Z statistic were positive or closer to zero than the “fixed  
13 critical value” then a statistician would conclude that the observed difference was  
14 not statistically significant and that there was no actual difference between the  
15 observed measurements.

16

17 Q. IF IT IS THAT SIMPLE TO USE A “FIXED CRITICAL VALUE” WHY DON’T  
18 WE JUST AGREE TO THAT APPROACH?

19

20 A. The problem is that while the “fixed critical value” can tell you whether the  
21 observed differences are statistically significant, it cannot tell you whether the  
22 differences are material. Let’s use an example. Suppose the observed interval for  
23 residential dispatched orders furnished to BellSouth’s retail operations is 4.1 days.  
24 Suppose the observed interval for the CLEC is 4.3 days. Using a “fixed critical  
25 value” it might be possible to get a Truncated Z statistic for these measurements

1 that was less than  $-1.645$ , that is, that was much larger in magnitude (farther from  
2 zero in the negative direction). That would tell you that the two numbers were  
3 statistically different. However, someone would then have to look at the actual  
4 numbers, 4.1 days versus 4.3 days, and determine whether the difference is  
5 material. Did it really make a difference to the CLEC or the CLEC's customers  
6 that it took two-tenths of a day longer, on average, to provide service to the  
7 CLEC's customer? Maybe it does and maybe it doesn't. Using the "fixed critical  
8 value" cannot answer that question, which means that another analysis will have  
9 to be made in each case where there is a statistically significant difference  
10 observed. This is not practical for a self-effectuating system that is supposed to  
11 determine parity on a timely basis.

12

13 Q. DOES THE USE OF THE "ERROR PROBABILITY BALANCING METHOD"  
14 FIX THIS PROBLEM?

15

16 A. It does. Using "Error Probability Balancing" we determine a "balancing critical  
17 value" which allows you to determine whether an observed difference is  
18 statistically significant and material all at the same time. Therefore, there is no  
19 need for another analysis and no dispute as to whether two-tenths of a day is  
20 material or not. The application of the "balancing critical value" provides both  
21 answers.

22

23 Q. CAN YOU TELL US MORE ABOUT THE DIFFERENCE BETWEEN THE  
24 "FIXED CRITICAL VALUE" AND THE "BALANCING CRITICAL VALUE?"

25

1 A. Certainly. I have already described how the “fixed critical value” is determined.  
2 The “balancing critical value” introduces another dimension and that involves  
3 what we call Type II errors. A Type II error is where the observed data suggests  
4 that parity has been achieved, but in fact it has not. In the simplest terms, a Type I  
5 error hurts the ILEC because it says the ILEC didn’t provide parity when in fact it  
6 did. A Type II error hurts the CLEC because it says that BellSouth provided  
7 parity when it did not. What the “Error Probability Balancing” method does is  
8 make the probability of committing either of the two different types of errors  
9 equal. You will recall when I was discussing the “fixed critical value” I talked  
10 only about having the probability of a Type I error at a level less than 5 percent.  
11 With a “balancing critical value,” we are saying that the number we are using to  
12 compare to the Z statistic reflects the probability that there will be just as many  
13 Type II errors as there are Type I errors. In other words, we don’t worry about  
14 whether there is a 5 percent chance of a Type I error or a 30 chance of a Type II  
15 error. Rather, we derive a figure that yields an equal probability of either type of  
16 error. There are formulae that are used to make the calculation that yields a single  
17 number that can be then compared to the Z statistic we talked about earlier.

18  
19 Q. CAN YOU DISCUSS THESE FORMULAE?

20  
21 A. The formulae are outlined in Appendix C of Exhibit EJM-1, and are difficult to  
22 describe in a short statement. The formulae are dependent upon the type of  
23 performance measure (mean, proportion, rate), the number of BellSouth and  
24 CLEC transactions, and the “delta” that is selected for use in the formula.

1 In a simple scenario with a large number of BellSouth transactions, an  
2 approximate value can be calculated by taking the negative of the square root of  
3 the number of CLEC transactions and multiplying it times the “delta” divided by  
4 2. I know that this is not intuitive, but once again these formulae are ones that a  
5 well-trained statistician would agree are appropriate, and would yield a critical  
6 value that represents a balancing of the Type I and Type II error probabilities. For  
7 instance, if we selected a “delta” of 1, and we had 25 CLEC observations, the  
8 appropriate critical value to compare the Truncated Z statistic to would be -2.5. If  
9 the Z statistic were less than -2.5 (that is, it is further from zero than -2.5) there  
10 would be a statistical difference and it would be material, thus avoiding the  
11 problems associated with the “fixed critical value” approach.

12  
13 If the Z statistic were greater than -2.5 (that is, the Z statistic was closer to zero or  
14 positive), it would indicate that the difference was not statistically significant and  
15 the analysis would be at an end.

16  
17 Q. WHY IS THIS METHODOLOGY APPROPRIATE?

18  
19 A. First of all, Dr. Mallows created the truncated Z statistic so that it possesses five  
20 important properties.

- 21  
22 1. It is a single, overall index on a standard scale; that is, you can use a normal  
23 bell shaped curve to make judgments.  
24 2. If transaction counts for BellSouth and the CLECs across comparison cells  
25 (classifications) are exactly proportional, the aggregate index should be very

nearly the same as if we had not disaggregated. This means that if the granular disaggregation I have discussed really wasn't necessary, you will still get the same results.

3. The contribution of each cell depends on the number of transactions in the cell.
4. As far as possible, systematic discriminatory performance in some cells is not masked by good performance in other cells.
5. The final result does not depend critically on minor details in the data; that is, small changes in transaction values only induce small changes in the final result.

Second, the methodology follows the four key principles that Dr. Mallows and the Ernst & Young team laid out.

1. Like-to-Like Comparisons. When possible, data should be compared at appropriate levels; for example, CLEC transactions that are "new" provisioning orders should be compared with "new" BellSouth provisioning orders.
2. Aggregate Level Test Statistic. Each performance measure of interest should be summarized by one overall test statistic giving the decision maker a rule that determines whether a statistically significant difference exists.
3. Production Mode Process. The decision system must be developed so that it does not require intermediate manual intervention.



1

2

3

4

5

6

7

Q. CAN YOU EXPLAIN WHAT THE TERM “DELTA” ENCOMPASSES?

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

4. Balancing. The testing methodology should balance Type I and Type II error probabilities. A Type I error adversely affects BellSouth; a Type II error adversely affects a CLEC. Balancing the error probabilities ensures that both sides assume the same level of uncertainty in the decision process.

A. “Delta” is a factor that is used to identify whether a meaningful difference exists between the BellSouth and CLEC performance, in addition to a statistically significant difference. It is a rather complex concept so let me try to use a very simple example to illustrate what “delta” does. I want to caution you that this is a simplistic example that I am offering just to try to illustrate this complex point. Let’s assume that within a particular wire center during the first half of a given month, the mean (average) time that BellSouth took to provision a new, dispatched, residential retail order with less than 10 circuits was 5 days. Assume further that the standard deviation associated with that mean or average was 1.5 days. This is a realistic example for what might be observed in a like-to-like cell, however it does not necessary represent what might happen across all comparison cells. Again, this example is being given to illustrate what “delta” does, and the specific values are not to be taken as representative of all possible situations.

If this were a normally distributed data set, then about 68 percent of all of these services were provisioned for BellSouth customers within a period of 3.5 days to 6.5 days. If we continued to assume a normally distributed data set, the remaining

1 32 percent of BellSouth's customers would fall equally above and below that  
2 spread of 3.5 to 6.5 days. Let's now assume that the "delta" or the materiality  
3 factor we choose was "1." This means that as long as the average time taken to  
4 provide the relevant service to the CLECs did not exceed the BellSouth mean (5  
5 days) plus one-half of the standard deviation I mentioned (1.5 days), the  
6 difference would not be material. That is, if the mean for the CLECs for this  
7 period were 5.75 days or less, the difference would not be discriminatory, and no  
8 penalty would be assessed. I arrived at the conclusion that the difference could  
9 not be more than one-half of the BellSouth standard deviation by dividing the  
10 "delta" of one by two, as I set out in my formula above. If the CLEC mean goes  
11 beyond 5.75 days, then BellSouth would be judge to be out of compliance, and  
12 assessed a penalty.

13  
14 Let's consider another very simple example to illustrate what happens when  
15 "delta" is reduced. Assume the exact same facts as above, but use a "delta" of  
16 "0.5." In that case, the difference between the BellSouth average for the month  
17 and the CLEC average for the month for the same measure could only be 5.375  
18 (five and three-eighths) days, instead of 5.75 (five and three-quarters) days. The  
19 question that the selection of "delta" raises is how close is close enough in terms  
20 of materiality. Is it material that BellSouth took three-quarters of a day longer, on  
21 average, to provide service to the CLEC than to its own retail services? Is it  
22 material that BellSouth took three-eighths of a day longer, on average?

23  
24 Q. HAVE THE STATISTICIANS DETERMINED THE APPROPRIATE VALUE  
25 FOR "DELTA"?

1

2 A. No. While statistical science can be used to evaluate the impact of different  
3 choices of these parameters, there is not much that an appeal to statistical  
4 principles can offer in directing specific choices. Specific choices should be made  
5 based on economic/business judgment.

6

7 Q. WHAT IS THE IMPACT OF THE AUTHORITY'S CHOICE OF A "DELTA" OF  
8 0.25 IN THE DELTACOM ARBITRATION?

9

10 A. By choosing a "delta" of 0.25, the Authority has decided that the appropriate  
11 reference alternative hypothesis to use in evaluating the probability of a Type II  
12 error is one where the CLEC mean is greater than the BellSouth by one-fourth of a  
13 standard deviation. Using the numbers from my previous example, the null  
14 hypothesis is that the CLEC mean is the same as the BellSouth mean of 5 days.  
15 The alternative hypothesis is that the CLEC mean is 5.375 days (5 and 3/8 days or  
16 5 days and 9 hours) when the BellSouth mean is 5 days. A balancing critical  
17 value for the statistical test is worked out so that the probability of concluding that  
18 CLEC customers receive the service in over 5 days on average when they actually  
19 receive service in 5 days, on average, is the same as the probability of concluding  
20 that CLEC customers receive service of 5 days on average, when they actually  
21 received service in 5.375 days on average.

22

23 But the story does not end there. If we look at how the test works in practice, we  
24 will see that when the observed difference in the CLEC-BellSouth average service  
25 is greater than one-eighth of a standard deviation, then BellSouth will be found to

1 be out of parity, and a penalty will be paid. So, if the BellSouth average service  
2 time is 5 days, on average, with a standard deviation of 1.5 days, then BellSouth  
3 will pay penalties whenever the average service time is greater than 5.1875 days  
4 (5 days plus 1.5 times one-half delta). Stating this example another way, if the  
5 CLEC average service time is more than 4 hours and 30 minutes longer than the  
6 BellSouth average of 5 days (4 hours, 30 minutes and 1 second or more) then  
7 BellSouth will be subjected to penalty payments.  
8

9 Q. DO ANY ASPECTS OF THE STATISTICAL METHODOLOGY NEED TO BE  
10 CHANGED FOR TIER II ENFORCEMENT MECHANISMS?  
11

12 A. No. The statistical methodology for comparing the service experience of all  
13 CLEC customers to BellSouth customers remains the same. One may want to  
14 consider changing the value of “delta” however. When the statisticians were  
15 putting together the “Statisticians’ Report” for Louisiana, it was thought that it  
16 might be prudent to use a smaller value of “delta” for Tier II testing. The  
17 reasoning behind this is that when one combines all CLEC transactions together,  
18 poor service to a few small CLEC’s could be masked by better service to the rest  
19 of the CLECs. One way to try to avoid such masking is to use a small materiality  
20 threshold. Whether or not this is necessary, and how much smaller “delta” should  
21 be for Tier II compared with Tier I, are questions subject matter experts and  
22 regulators should answer. As was stated before, the statistician should still play a  
23 role in this process so that the impact of various choices can be assessed.  
24

1 Q. WOULD YOU PLEASE EXPLAIN THE CONCEPT BEHIND THE VOLUME  
2 PROPORTION CALCULATION THAT IS USED TO DETERMINE THE  
3 PENALTY PAYMENT ORDERED BY THE AUTHORITY IN THE  
4 DELTACOM ARBITRATION AND ADVOCATED BY BELL SOUTH?

5  
6 A. Certainly. It is my understanding that BellSouth wants to have a penalty plan  
7 based on the number of CLEC transactions that, if changed for the better, would  
8 change a Truncated Z test failure into a pass. That is, under BellSouth's plan, a  
9 calculation is made of the number of transactions that would have had to be  
10 accomplished more quickly (if the time interval was the relevant measure) in  
11 order to avoid having a failure. The penalty is then determined by multiplying  
12 that number of transactions times the appropriate penalty amount. This concept is  
13 used in Southwestern Bell's Texas penalty plan. The problem for BellSouth is  
14 that the Truncated Z statistic is much more complex than the statistics that  
15 Southwestern Bell uses, and this makes it very difficult to directly observe or  
16 calculate the number of transactions in any particular situation which would have  
17 to be changed to turn a failure into a pass. That being the case, BellSouth  
18 developed a surrogate model, which I discuss next, for calculating the number of  
19 transactions subject to remedy.

20  
21 When BellSouth fails a test, that is the Truncated Z statistic is found to be less  
22 than the critical value (further from zero in the negative direction), the calculation  
23 of the penalty amount proceeds as follows:  
24

- 1       1. Determine the “impacted volume,” that is the number of CLEC transactions in  
2       cells with negative z-values that are potentially subjected to penalties. These  
3       are transactions that, if changed for the better, would increase a Truncated Z  
4       value in the positive direction.  
5
- 6       2. Calculate the “parity gap.” Subtract the Truncated Z statistic from the  
7       balancing critical value. For example, if the Truncated Z statistic for a  
8       particular submeasure is calculated to be –3, and the balancing critical value is  
9       –2, then there is a test failure. The “parity gap” is the distance between –2 and  
10      –3, which is 1 unit. If the Truncated Z statistic turned out to be –4 instead,  
11      then the “parity gap” would be 2 since the distance between –4 and –2 is two  
12      units.  
13
- 14     3. Calculate the volume proportion. Divide the “parity gap” by 4. If the “parity  
15      gap” is larger than 4, than the volume proportion is capped at 100%.  
16      Otherwise, the “volume proportion” will be between 0 and 100%. For  
17      example, when the “parity gap” is 1, then the “volume proportion” is 1  
18      divided by 4, or 25%. If the “parity gap” is 2, then the “volume proportion” is  
19      50%. The volume proportion keeps growing, as the gap gets larger. Once the  
20      gap reaches 4 units, the volume proportion reaches 100%. Since you could  
21      not change any more impacted transaction for the better, we stop here, and set  
22      the volume proportion to 100% for all gaps larger than 4 units.  
23
- 24     4. Calculate “effected volume” by multiplying the “impacted volume” by the  
25      “volume proportion.” This is the number of transaction for which a penalty

1 will be paid. The “effected volume” is the surrogate model’s estimate for the  
2 number of transactions to change in order to get the Truncated Z value equal  
3 to the balancing critical value. For example, if the “impacted volume” is 100  
4 CLEC transactions, and the “volume proportion” is 50%, then the “effected  
5 volume” is 50 transactions.

- 6
- 7 5. Calculate the final penalty payment. Multiply the “effected volume” by the  
8 per transaction fee of the submeasure class (given by BellSouth witness Mr.  
9 Coon in Exhibit No. DAC-2).

10

11 Q. WHAT DO YOU MEAN BY TRANSACTIONS THAT ARE “CHANGED FOR  
12 THE BETTER?”

13

14 A. The term “change for the better” can mean different things for different types of  
15 measures. For a proportion measure, such as missed installations, “poor”  
16 transactions are those that are not completed on time (they are “missed”).  
17 Changing a “poor” transaction for the better would amount to changing a  
18 “missed” transaction into a “non-miss,” or completed on time. So for a proportion  
19 measure we would want to start to define the impacted volume with all “missed”  
20 CLEC transactions. If we are dealing with the rate measure customer trouble  
21 report rate, a “poor” transaction is any trouble, and if the trouble never occurred  
22 that would be “better.” So for this rate measure, defining the impacted volume  
23 starts with all CLEC troubles. For a metric that is measuring the average duration  
24 of a service, such as order completion interval, a change for the better means that  
25 the service time is made shorter. This could be any CLEC service order, so for

1 mean measures we would want to start to define the impacted volume with all  
2 CLEC transactions.

3

4 Q. WHY ARE TRANSACTIONS FROM CELLS WITH NEGATIVE Z-VALUES  
5 THE ONLY ONES USED TO DEFINE IMPACTED VOLUME?

6

7 A. Only transactions from negative cells that are “changed for the better” can  
8 increase the Truncated Z value in the positive direction. If we used all CLEC  
9 transactions that could be changed for the better to define the impacted volume,  
10 we would be including transactions that will not increase the Truncated Z value  
11 even if the transactions are “changed for the better.” Recall, that in order to  
12 compute the Truncated Z statistic, we disaggregate the data into many like-to-like  
13 cells, and computed cell z-values. In calculating the value of the Truncated Z,  
14 positive cell z-values are set to zero. These cells do contribute to the final value  
15 of the Truncated Z, but changing “poor” transactions to “better” values in such  
16 cells, will not change how much these cells contributed to the Truncated Z. This  
17 is because these cells already exhibit service that is favorable to the CLEC.  
18 Changing transactions for the better will still result in a positive cell z-value  
19 which is still set to zero for calculating the Truncated Z. Our goal is to increase  
20 the value of the Truncated Z result by changing transactions for the better. This  
21 will not happen for cells with positive z-values. Thus, only CLEC transactions in  
22 cells with negative z-values can be changed for the better, and improve the value  
23 of the Truncated Z score. So, the impacted volume is the number of CLEC  
24 transactions in “negative” cells that can be changed for the better.

25



1 Q. IS THERE A WAY TO CALCULATE THE NUMBER OF TRANSACTIONS  
2 WITHOUT THE USE OF A SURROGATE MODEL?

3

4 A. Yes, if you have a proportion or rate measure. It turns out that a mathematical  
5 solution, called a linear program, can be used to find the maximum number of  
6 “missed” transactions in negatively effected cells that, if changed to “non-misses,”  
7 will shrink the gap between the balancing critical value and the Truncated Z  
8 statistic to zero. Linear programs can be very time consuming to solve, especially  
9 if the Truncated Z test involves a large number of cells. Thus, one may still want  
10 to use a surrogate model to be able to calculate the transaction count. Even if a  
11 linear program is used for proportion and rate measures, it will not work for a  
12 mean measure. Using the same surrogate model for all measure types seems like  
13 a reasonable thing to do.

14

15 Q. HAVE YOU EVER USED THE LINEAR PROGRAMMING METHOD TO  
16 CALCULATE THE NUMBER OF TRANSACTIONS FOR THE PENALTY  
17 CALCULATION AND COMPARED THAT TO THE RESULTS OBTAINED BY  
18 USING BELL SOUTH’S SURROGATE METHOD?

19

20 A. Yes. The Ernst & Young statistical team examined Louisiana data from  
21 November 1999 for the percent missed installations performance measurement.  
22 We were able to solve the linear program for 39 Tier I and Tier II tests in which  
23 BellSouth failed a parity test.

24

25 Q. WHAT WERE THE RESULTS OF YOUR ANALYSIS?

1

2 A. For each of the 39 tests, BellSouth's SEEM plan (the surrogate model) calculated  
3 the same number, or a higher number, of transactions for penalty payment than the  
4 linear program solution. In fact, the linear program calculated a total of 352  
5 transactions for penalties across the 39 tests. The SEEM plan, on the other hand,  
6 calculated a total of 814 transactions for penalties. This means that the SEEM  
7 plan was calculating more transactions to "change for the better" than needed for  
8 the purpose of shrinking the gap between the Truncated Z statistic and the  
9 balancing critical value.

10

11 Q. WOULD YOU SUMMARIZE THE KEY ISSUES THAT THE AUTHORITY  
12 NEEDS TO CONSIDER IN ADOPTING THE METHODOLOGY YOU ARE  
13 RECOMMENDING?

14

15 A. Yes. In order to carry out the Truncated Z with Error Probability Balancing, the  
16 Authority needs to evaluate two key aspects of any proposed plan: 1) the level of  
17 aggregation at which parity decisions will be made, and 2) the "delta" value used  
18 to determine the balancing critical value. Neither of these input parameters is  
19 something that should be decided upon solely by statisticians. Input from subject  
20 matter experts is needed.

21

22 For the aggregation level, I believe that BellSouth is proposing that the same level  
23 of aggregation that the Authority ordered in the Deltacom arbitration be used for  
24 the penalty plan. On the other hand, BellSouth is recommending a different  
25 "delta" value than the one in the Deltacom arbitration. While I do not have a

1           specific recommendation for the “delta” value, I believe that it should be  
2           understood that the impact of any choice of “delta” is that BellSouth will pay  
3           penalties when the observed difference in the means of the ILEC and CLEC is  
4           greater than one-half “delta” standard deviation, as described in the examples I  
5           have already given.

6

7    Q.     DOES THIS CONCLUDE YOUR TESTIMONY?

8

9    A.     Yes.

AFFIDAVIT

DISTRICT OF COLUMBIA

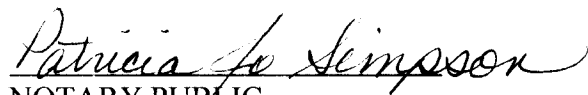
BEFORE ME, the undersigned authority, duly commissioned and qualified in and for the District aforesaid, personally came and appeared Edward J. Mulrow –Senior Manager – Quantitative Economics and Statistics Group, Ernest & Young, LLP, who, being by me first duly sworn deposed and said that:

He is appearing as a witness before the Tennessee Regulatory Authority in Docket No. 01-00193 on behalf of BellSouth Telecommunications, Inc., and if present before the Authority and duly sworn, his testimony would be set forth in the annexed testimony consisting of 24 pages and 2 exhibit(s).



Edward J. Mulrow

Sworn to and subscribed  
before me on JUNE 5, 2001



NOTARY PUBLIC

DISTRICT OF COLUMBIA

My COMMISSION EXPIRES MARCH 31, 2006



BellSouth Telecommunications, Inc. 504 528-2050  
Suite 3060 Fax 504 528-2948  
365 Canal Street  
New Orleans, Louisiana 70130-1102

Victoria K. McHenry  
General Counsel - LA

February 29, 2000

**VIA FEDERAL EXPRESS**

Ms. Susan Cowart  
Louisiana Public Service Commission  
Suite 1630  
One American Place  
Baton Rouge, LA 70825

RE: I.PSC Docket No. U-22252-C  
Louisiana Public Service Commission, ex parte  
In re: BellSouth Telecommunications, Inc.  
Service Quality Performance Measurements

Dear Ms. Cowart:

Enclosed are the original and one (1) copy each of the following documents to be filed into the record of the referenced matter:

1. Updated BellSouth SQM Report
2. Updated Statistician's Report

These items were not specifically included in the Commission's most recently issued Notice so I am unsure when they are due. In any event, we are providing them as soon as possible.

Sincerely,

A handwritten signature in cursive script that reads 'Victoria K. McHenry'. Below the signature is a horizontal line with the number '98' written at the end.

Victoria K. McHenry

VKM/as  
Encs.

cc: Official Service List (w/enc.)(via email)

DRAFT

## Statistical Techniques For The Analysis And Comparison Of Performance Measurement Data

Submitted to Louisiana Public Service Commission (LPSC)  
Docket U-22252 Subdocket C

Revised February 28, 2000

### Introduction and Scope

The Louisiana Public Service Commission (LPSC) staff has requested Drs. S. Hinkins, E. Mulrow, and F. Scheuren<sup>1</sup> of Ernst & Young LLP (consultants for BellSouth Telecommunications), and Dr. C. Mallows of AT&T Labs-Research to set out their views on the application of a statistical analysis to performance measurement data. The present report is intended to provide a detailed statistical report on appropriate methodology.

The setting for the analysis is crucial to the interpretation of any statistical significance that might be found. There is no doubt that, to quote the Commission staff, "statistical analysis can help reveal the likelihood that reported differences in an ILECs performance toward its retail customers and CLECs are due to underlying differences in behavior rather than random chance" (Staff Final Recommendation, LPSC Docket No. U-22252 - Subdocket C, dated August 12, 1998, pages 15 - 16).

To frame our presentation the next paragraph from the LPSC Docket U-22252 is quoted in its entirety.

**"Statistical tests are effective in identifying those measurements where differences in performance exist. The tests themselves cannot identify the cause of the apparent differences. The differences may be due to a variety of reasons, including: 1) when the ILEC and CLEC processes being measured are actually different and should not be expected to produce the same result, 2) when the ILEC is employing discriminatory practices, or 3) when assumptions necessary for the statistical test to be valid are not being met." (Ibid., page 16)**

Apparent statistically significant differences in BellSouth and CLEC performance can arise when

- the ILEC and CLEC processes being measured are actually different and should not be expected to produce the same result
- the ILEC is employing discriminatory practices, or
- assumptions necessary for the statistical test to be valid are not being met.

<sup>1</sup> Dr. Scheuren is now a Senior Fellow at the Urban Institute.

RECEIVED

MAR 01 2000

LOUISIANA PUBLIC SERVICE COMMISSION  
ADMINISTRATIVE HEARINGS DIVISION

## DRAFT

To meet the Louisiana Commission's purpose, we will recommend techniques that are robust in the presence of possible assumption failure, carefully examine BellSouth Telecommunications (BST) and CLEC performance so "like" is compared only to "like," and are still able, in a highly efficient manner, to detect differences. Upon investigation any differences detected might lead to concerns about possible discriminatory practices.

The LPSC staff also states "that a uniform methodology which identifies those items which need to be measured, how they are to be measured, and how the results are to be reported is also desirable and would be beneficial to all parties" (*Ibid.*, page 16). We agree with this goal as well, stipulating only that the use of a single method may not be desirable while a single methodology (or a set of methods) could be.

The statistical process for testing if CLEC and ILEC customers are being treated equally involves more than just a mathematical formula. Three key elements need to be considered before an appropriate decision process can be developed. These are

- the type of data,
- the type of comparison, and
- the type of performance measure.

When examining the various combinations of these elements, we find that there is a set of testing principles that can be applied uniformly. However, the statistical formulae that need to be used change as the situation changes.

To be responsive to the Commission, we have divided our discussion into four sections and five appendices. The contents of each of these are briefly mentioned below -- first for the main report and then for the extensive supporting appendix materials.

For the main report, this section (Section I) introduces our work and sets out the required scope. The next two sections (Sections II and III) discuss the type of comparisons that need to be identified, and the appropriate testing principles. The final section (Section IV) provides an overview of appropriate testing methodologies, based on what we have learned from our examination of BellSouth's performance measure data in Louisiana.

The five appendices provide technical details on the statistical calculations involved in the Truncated Z statistic (Appendix A), the implementation of the methodology for the trunk blocking performance measure (Appendix B), the calculations involved in computing the balancing critical value of a test (Appendix C), examples of ways to present the results using detailed statistical displays so that results can be audited (Appendix D), and the technical details involved in data trimming (Appendix E).

## DRAFT

### 2. Data Considerations, Comparisons, and Measurement Types

This section makes general distinctions which apply to the performance measures. These distinctions will be important in the determination of appropriate methodologies.

**Data Set Types.** The type of statistical methodology used depends on the form of the data available. In general, there are two ways to classify the data used for performance measure comparisons. These are:

- transaction level data, and
- aggregated summaries.

Records in a transaction level data set represent a single transaction, e.g. an individual customer order, or the record of a specific trouble reported by a customer. This type of data set allows for deep like-to-like comparisons, and may also allow one to identify the root cause of a problem. A testing methodology needs to be carefully chosen so that it incorporates the comparison levels and does not cover up problem areas.

Records in an aggregated summary data set are typically summaries of related transactions. For example, the total number of blocked calls in a trunk group during the noon hour of a day is a summary statistic. This type of data set may not contain as much information as a transaction level data set, and it therefore needs to be treated differently. While a general methodology may be determined for a transaction level data set, it may not be possible to do so for aggregated summaries. Testing methodology needs to be developed on a case-by-case basis.

**Comparison Types.** An ILEC's performance in providing services to CLEC customers is tested in one of two ways:

- by comparing CLEC performance to ILEC performance when a retail analog exists, or
- by comparing CLEC performance to a benchmark.

The testing methodologies for these two situations will have similarities, but there are differences that need to be understood.

Table 1 categorizes those performance measures that E&Y has examined by data type and comparison type. The table shows that five performance measures with retail analogs have transaction level data, while three others with retail analogs only have summary level data. No performance measures using benchmarks have been studied.



## DRAFT

**Table 1. Classification of Performance Measures by Data and Comparison Type**  
(only measures previously examined by E&Y are included)

Level of Data	Comparison Type	
	Retail Analog	Benchmark
Transaction Level	Order Completion Interval	No Measures Examined
	Maintenance Average Duration	
	% Missed Installations	
	% Missed Repair	
	Trouble Report Rate	
Summary Level	Billing Timeliness	No Measures Examined
	OSS Response Interval	
	Trunk Blocking	

**Measurement Types.** The performance measures that will undergo testing are of four types: means, proportions (an average of a measure that takes on only the values of 0 or 1), rates, and ratios.

While all four have similar characteristics, proportions and rates are derived from count data while means and ratios are derived from interval measurements. Table 2 classifies the performance measures by the type of measurement.

**Table 2: Classification of Performance Measures by Measurement Type**

Mean	Proportion	Rate	Ratio
Order Completion Interval Maint. Ave. Duration OSS Response Interval	Percent Missed Installations Percent Missed Repairs Billing Timeliness Trunk Blocking	Trouble Report Rate	Billing Accuracy

### 3. Testing Principles

This section describes five general principles which the final methodology should satisfy:

## DRAFT

1. *When possible, data should be compared at appropriate levels, e.g. wire center, time of month, dispatched, residential, new orders.*
2. *Each performance measure of interest should be summarized by one overall test statistic giving the decision maker a rule that determines whether a statistically significant difference exists.*
3. *The decision system must be developed so that it does not require intermediate manual intervention.*
4. *The testing methodology should balance Type I and Type II Error probabilities.*
5. *Trimming of extreme observations from BellSouth and CLEC distributions is needed in order to ensure that a fair comparison is made between performance measures.*

Like-to-Like Comparisons. *When possible, data should be compared at appropriate levels, e.g. wire center, time of month, dispatched, residential, new orders.*

In particular, to meet this goal the testing process should:

- Identify variables that may affect the performance measure.
- Record important confounding covariates.
- Adjust for the observed covariates in order to remove potential biases and to make the CLEC and the ILEC units as comparable as possible.

It is a well known principle that comparisons should be made on equal footing: apples-to-apples, oranges-to-oranges. Statistical techniques that are addressed in most text books usually assume that this is the case beforehand. Some higher level books address the issue of “designed experiments” and discuss appropriate ways to structure the data collection method so that the text books’ formulae can be used in analyzing the data.

Performance measure testing does not involve data from a designed experiment. Rather, the data is obtained from an observational study. That being the case, one must impose a structure on the data after it is gathered in order to assure that fair comparisons are being made. For example, it is important to disaggregate the data to a fine level so that appropriate like-to-like comparisons of CLEC and ILEC data can be made. Any statistical methodology that ignores important confounding variables can produce biased results.

Aggregate Level Test Statistic. *Each performance measure of interest should be summarized by one overall test statistic giving the decision maker a rule that determines whether a statistically significant difference exists.*

To achieve this goal, the aggregate test statistic should have the following properties:

## DRAFT

- The method should provide a single overall index, on a standard scale.
- If entries in comparison cells are exactly proportional over a covariate, the aggregated index should be very nearly the same as if comparisons on the covariate had not been done.
- The contribution of each comparison cell should depend on the number of observations in the cell.
- Cancellation between comparison cells should be limited, i.e., positive outcomes should not be allowed to cancel negative ones.
- The index should be a continuous function of the observations.

Since the data are being disaggregated to a very deep level, thousands of like-to-like comparison cells are created. An aggregate summary statistic is needed in order to make an overall judgment.

The aggregate level statistic should be insensitive to small changes in cells values, and its value should not be affected if some of the disaggregation for like-to-like cells is truly unnecessary. Furthermore, individual cell results should be weighted so that those cells with more transactions have larger effects on the overall result.

**Production Mode Process.** *The decision system must be developed so that it does not require intermediate manual intervention.*

Two statistical paradigms are possible for examining performance measure data. In the exploratory paradigm, data are examined and methodology is developed that is consistent with what is found. In a production paradigm a methodology is decided upon before data exploration. For the production paradigm to succeed

- Calculations should be well defined for possible eventualities.
- The decision process should be based on an algorithm that needs no manual intervention.
- Results should be arrived at in a timely manner.
- The system must recognize that resources are needed for other performance measure-related processes that also must be run in a timely manner.
- The system should be both auditable and adjustable over time.

While the exploratory paradigm provides protection against using erroneous data, it requires a great deal of lead time and is unsuitable for timely monthly performance measure testing. A production paradigm will not only promptly produce overall test results but will also provide documentation that can be used to explore the data after the test results are released.

## DRAFT

Error Probability Balancing. *The testing methodology should balance Type I and Type II Error probabilities.*

Specifically, what is required to achieve this goal is

- The probability of a Type I error should equal the probability of a Type II error for well-defined null and alternative hypotheses.
- The formula for a test's balancing critical value should be simple enough to calculate using standard mathematical functions, i.e. one should avoid methods that require computationally intensive techniques.
- Little to no information beyond the null hypothesis, the alternative hypothesis, and the number of observations should be required for calculating the balancing critical value.

The objective of a statistical test is to test a hypothesis concerning the values of one or more population parameters. Usually an inquiry into whether or not there is evidence to support a hypothesis, called the *alternative hypothesis*, is conducted by seeking statistical evidence that the converse of the alternative, the *null hypothesis*, is most likely false. If there is not sufficient evidence to reject the null hypothesis, then a case for accepting the alternative has not been made.

Two types of errors are possible in any decision-making process. These have been summarized in Table 3.

**Table 3: Statistical Testing Errors**

Decision Error	General Description	In terms of Performance Measure Testing
Type I	Rejecting the null hypothesis (accepting the alternative) when the null is true.	Deciding that BST favors its own customers when it does not.
Type II	Accepting the null hypothesis when the alternative is true.	Deciding that BST does not favor its own customers when it does.

In a controlled experimental study where the sample sizes are relatively small, it is generally desirable to control the Type I error closely to avoid making a conclusion that there is a difference when, in fact, there is none. The probability of a Type II error is not directly controlled but is determined by the sample size and the distance between the null and the alternative hypotheses.

## DRAFT

If a standard of materiality is set by stating a specific alternative for the test, and the distribution of the test statistic under both the null and alternative hypotheses is understood, then a critical value can be determined so that the two error probabilities are equal.

Trimming. *Trimming of extreme observations from BellSouth and CLEC distributions is needed in order to ensure that a fair comparison is made between performance measures.*

Three conditions are needed to accomplish this goal. These are:

- Trimming should be based on a general rule that can be used in a production setting.
- Trimmed observations should not simply be discarded; they need to be examined and possibly used in the final decision making process.
- Trimming should only be used on performance measures that are sensitive to "outliers."

For the purpose of performance measure testing, trimming refers to removing transactions that significantly distort the performance measure statistic for the set of transactions under consideration. For example, the arithmetic average (or mean) is extremely sensitive to "outliers" since a single large value can significantly distort the average.

The term "outliers" refers to:

- 1) extreme data values that may be valid, but since they are rare measurements, they may be considered to be statistically unique; or
- 2) large values that should not be in the analysis data set because of errors in the measurement or in selecting the data.

Trimming is beneficial since it puts both ILEC and CLEC transactions on equal footing with respect to the largest value in each set. Note, though, that it is only needed for performance measures that are distorted by outliers. Of the three types of measures defined in Section 2, only mean (average) measures require trimming. Appendix E sets forth a trimming plan for mean performance measures.

#### **4. Testing Methodology**

This section details the testing methodology that is most appropriate for the various types of performance measures. First, transaction level testing will be discussed when there is a retail analog. Next, transaction level testing against a benchmark. Then, testing when only aggregated summaries are available.

Transaction Level - Retail Analog: The Truncated Z Statistic. When a retail analog is available CLEC performance can be directly compared with ILEC performance. Over

## DRAFT

the last year. for transaction level data, many test statistics have been examined. We now believe that the "Truncated Z" test statistic provides the best compromise with respect to possessing the desired qualities outlined in Section 3, above.

The Truncated Z is fully described in Appendix A, and formulae for calculation of a balancing critical value are found in Appendix C. The main features of this statistic are:

- A basic test statistic is calculated within each comparison cell.
- The value of a cell's result is left "as is" if the result suggests that "favoritism" may be taking place. Otherwise, the result is set to zero. This is called the truncation step.
- Weights that depend on the volume of both ILEC and CLEC transactions within the cell are determined, and a weighted sum of the "truncated" cell results is calculated.
- The weighted sum is theoretically corrected to account for the truncation, and a final overall statistic is determined.
- This overall test value is compared to a balancing critical value to determine if favoritism is likely.

The test statistic itself is based on like-to-like comparisons, and it possesses all five of the properties of an aggregate test statistic (Section 3). While the test requires a large amount of calculations, our studies of the process on some of BellSouth's performance measure data indicate that the calculations can be completed in a reasonable amount of time. Therefore, the process can be put into production mode. Finally, since a balancing critical value can be calculated, it is possible to balance the error probabilities.

Transaction Level - Benchmark. When a benchmark is used, CLEC performance is not compared with ILEC performance. Like-to-like comparison cells are not needed, thus greatly simplifying the testing process. Statistical testing can be done using a probability model, or non-statistical testing can be done using a deterministic model. No data for this data/comparison class has been studied at this point in time.

Aggregated Summary - Retail Analog or Benchmark. We cannot provide any one single set of rules for the analysis of data in this class. Data that is an aggregated summary of transactions may or may not present problems. For example, BellSouth's trunk blocking data is saved as summaries by hour of the day. Collectively, the summaries do provide sufficient information to proceed with the Truncated Z methodology.

On the other hand, our examination of the data for the OSS response interval revealed that information necessary for computing a Truncated Z was not available. In this case, however, we were able to construct a satisfactory time series method to analyze the measure.

## **DRAFT**

Each measure falling into this class needs to be handled on a case-by-case basis. If sufficient information is available to use the Truncated Z method, then we feel it should be used. When the Truncated Z cannot be used, a testing methodology that adheres closely to the principles outlined in Section 3 should be determined and followed.

## Appendix A. The Truncated Z Statistic

The Truncated Z test statistic was developed by Dr. Mallows in order to have an aggregate level test when transaction level data are available that

- provides a single overall index on a standard scale;
- will not change the outcome if the disaggregation is unnecessary,
- incorporates the number of observations in a cell into the determination of the weight for the contribution of each comparison cell,
- limits the amount of “neutralization” between comparison cells, and
- is a continuous function of the observations.

The Ernst & Young statistical team and Dr. Mallows have studied the implementation of the statistic using some of BellSouth’s performance measure data. This has resulted in an overall process for comparing CLEC and ILEC performance such that the following principles hold:

- 1) Like-to-Like Comparisons are made. (See Appendix B for an example based on the trunk blocking measure.)
- 2) Error probabilities are balanced. (See Appendix C)
- 3) Extreme values are trimmed from the data sets when they significantly distort the performance measure statistic. (See Appendix E)
- 4) The testing process is an automated production system. (Discussed here. See Appendix D for reporting guidelines.)
- 5) The determination of ILEC favoritism is based on a single aggregate level test statistic. (Discussed here.)

This appendix provides the details behind computing the Truncated Z test statistic so that principles 4 and 5 hold. We start by assuming that any necessary trimming of the data is complete, and that the data are disaggregated so that comparisons are made within appropriate classes or adjustment cells that define “like” observations.

### Notation and Exact Testing Distributions

Below, we have detailed the basic notation for the construction of the truncated z statistic. In what follows the word “cell” should be taken to mean a like-to-like comparison cell that has both one (or more) ILEC observation and one (or more) CLEC observation.

- $L$  = the total number of occupied cells
- $j$  =  $1, \dots, L$ ; an index for the cells
- $n_{1j}$  = the number of ILEC transactions in cell  $j$
- $n_{2j}$  = the number of CLEC transactions in cell  $j$
- $n_j$  = the total number transactions in cell  $j$ ;  $n_{1j} + n_{2j}$



$$\begin{aligned}
X_{1jk} &= \text{individual ILEC transactions in cell } j; k = 1, \dots, n_{1j} \\
X_{2jk} &= \text{individual CLEC transactions in cell } j; k = 1, \dots, n_{2j} \\
Y_{jk} &= \text{individual transaction (both ILEC and CLEC) in cell } j \\
&= \begin{cases} X_{1jk} & k = 1, K, n_{1j} \\ X_{2jk} & k = n_{1j} + 1, K, n_j \end{cases}
\end{aligned}$$

$\Phi^{-1}(\cdot)$  = the inverse of the cumulative standard normal distribution function

For Mean Performance Measures the following additional notation is needed.

$$\begin{aligned}
\bar{X}_{1j} &= \text{the ILEC sample mean of cell } j \\
\bar{X}_{2j} &= \text{the CLEC sample mean of cell } j \\
s_{1j}^2 &= \text{the ILEC sample variance in cell } j \\
s_{2j}^2 &= \text{the CLEC sample variance in cell } j \\
\{y_{jk}\} &= \text{a random sample of size } n_{2j} \text{ from the set of } Y_{j1}, K, Y_{jn_j}; k = 1, \dots, n_{2j} \\
M_j &= \text{the total number of distinct pairs of samples of size } n_{1j} \text{ and } n_{2j}; \\
&= \binom{n_j}{n_{1j}}
\end{aligned}$$

The exact parity test is the permutation test based on the "modified Z" statistic. For large samples, we can avoid permutation calculations since this statistic will be normal (or Student's t) to a good approximation. For small samples, where we cannot avoid permutation calculations, we have found that the difference between "modified Z" and the textbook "pooled Z" is negligible. We therefore propose to use the permutation test based on pooled Z for small samples. This decision speeds up the permutation computations considerably, because for each permutation we need only compute the sum of the CLEC sample values, and not the pooled statistic itself.

A permutation probability mass function distribution for cell j, based on the "pooled Z" can be written as

$$PM(t) = P\left(\sum_k y_{jk} = t\right) = \frac{\text{the number of samples that sum to } t}{M_j},$$

and the corresponding cumulative permutation distribution is

$$\text{CPM}(t) = P\left(\sum_k y_{jk} \leq t\right) = \frac{\text{the number of samples with sum} \leq t}{M_j}.$$

For Proportion Performance Measures the following notation is defined

- $a_{1j}$  = the number of ILEC cases possessing an attribute of interest in cell  $j$
- $a_{2j}$  = the number of CLEC cases possessing an attribute of interest in cell  $j$
- $a_j$  = the number of cases possessing an attribute of interest in cell  $j$ ;  $a_{1j} + a_{2j}$

The exact distribution for a parity test is the hypergeometric distribution. The hypergeometric probability mass function distribution for cell  $j$  is

$$\text{HG}(h) = P(H = h) = \begin{cases} \frac{\binom{n_{1j}}{h} \binom{n_{2j}}{a_j - h}}{\binom{n_j}{a_j}}, & \max(0, a_j - n_{2j}) \leq h \leq \min(a_j, n_{1j}) \\ 0 & \text{otherwise} \end{cases},$$

and the cumulative hypergeometric distribution is

$$\text{CHG}(x) = P(H \leq x) = \begin{cases} 0 & x < \max(0, a_j - n_{2j}) \\ \sum_{h=\max(0, a_j - n_{2j})}^x \text{HG}(h), & \max(0, a_j - n_{2j}) \leq x \leq \min(a_j, n_{1j}) \\ 1 & x > \min(a_j, n_{1j}) \end{cases}.$$

For Rate Measures, the notation needed is defined as

- $b_{1j}$  = the number of ILEC base elements in cell  $j$
- $b_{2j}$  = the number of CLEC base elements in cell  $j$
- $b_j$  = the total number of base elements in cell  $j$ ;  $b_{1j} + b_{2j}$
- $\bar{p}_{1j}$  = the ILEC sample rate of cell  $j$ ;  $n_{1j}/b_{1j}$
- $\bar{p}_{2j}$  = the CLEC sample rate of cell  $j$ ;  $n_{2j}/b_{2j}$
- $q_j$  = the relative proportion of ILEC elements for cell  $j$ ;  $b_{1j}/b_j$

The exact distribution for a parity test is the binomial distribution. The binomial probability mass function distribution for cell  $j$  is

$$BN(\cdot) = P(B = k) = \begin{cases} \binom{n_j}{k} q_j^k (1 - q_j)^{n_j - k}, & 0 \leq k \leq n_j, \\ 0 & \text{otherwise} \end{cases}$$

and the cumulative binomial distribution is

$$CBN(x) = P(B \leq x) = \begin{cases} 0 & x < 0 \\ \sum_{k=0}^x BN(k), & 0 \leq x \leq n_j. \\ 1 & x > n_j \end{cases}$$

For Ratio Performance Measures the following additional notation is needed.

- $U_{1jk}$  = additional quantity of interest of an individual ILEC transaction in cell  $j$ ;  $k = 1, \dots, n_{1j}$
- $U_{2jk}$  = additional quantity of interest of an individual CLEC transaction in cell  $j$ ;  $k = 1, \dots, n_{2j}$
- $\hat{R}_{ij}$  = the ILEC ( $i = 1$ ) or CLEC ( $i = 2$ ) ratio of the total additional quantity of interest to the base transaction total in cell  $j$ , i.e.,  $\sum_k U_{ijk} / \sum_k X_{ijk}$

### Calculating the Truncated Z

The general methodology for calculating an aggregate level test statistic is outlined below.

1. **Calculate cell weights,  $W_j$ .** A weight based on the number of transactions is used so that a cell which has a larger number of transactions has a larger weight. The actual weight formulae will depend on the type of measure.

*Mean or Ratio Measure*

$$W_j = \sqrt{\frac{n_{1j}n_{2j}}{n_j}}$$

*Proportion Measure*

$$W_j = \sqrt{\frac{n_{2j}n_{1j}}{n_j} \cdot \frac{a_j}{n_j} \cdot \left(1 - \frac{a_j}{n_j}\right)}$$

*Rate Measure:*

$$W_j = \sqrt{\frac{b_{1j}b_{2j}}{b_j} \cdot \frac{n_j}{b_j}}$$

2. **In each cell, calculate a Z value,  $Z_j$ .** A Z statistic with mean 0 and variance 1 is needed for each cell.

- If  $W_j = 0$ , set  $Z_j = 0$ .
- Otherwise, the actual Z statistic calculation depends on the type of performance measure.

*Mean Measure*

$$Z_j = \Phi^{-1}(\alpha)$$

where  $\alpha$  is determine by the following algorithm.

If  $\min(n_{1j}, n_{2j}) > 6$ , then determine  $\alpha$  as

$$\alpha = P(t_{n_{1j}-1} \leq T_j),$$

that is,  $\alpha$  is the probability that a t random variable with  $n_{1j} - 1$  degrees of freedom, is less than

$$T_j = t_j + \frac{g}{6} \left( \frac{n_{1j} + 2n_{2j}}{\sqrt{n_{1j} n_{2j} (n_{1j} + n_{2j})}} \right) \left( t_j^2 + \frac{n_{2j} - n_{1j}}{2n_{1j} + n_{2j}} \right),$$

where

$$t_j = \frac{\bar{X}_{1j} - \bar{X}_{2j}}{s_{1j} \sqrt{\frac{1}{n_{1j}} + \frac{1}{n_{2j}}}}$$

and the coefficient  $g$  is an estimate of the skewness of the parent population, which we assume is the same in all cells. It can be estimated from the ILEC values in the largest cells. This needs to be done only once for each measure. We have found that attempting to estimate this skewness parameter for each cell separately leads to excessive variability in the "adjusted" t. We therefore use a single compromise value in all cells.

Note, that  $t_j$  is the "modified Z" statistic. The statistic  $T_j$  is a "modified Z" corrected for the skewness of the ILEC data.

If  $\min(n_{1j}, n_{2j}) \leq 6$ , and

a)  $M_j \leq 1,000$  (the total number of distinct pairs of samples of size  $n_{1j}$  and  $n_{2j}$  is 1,000 or less).

- Calculate the sample sum for all possible samples of size  $n_{2j}$ .
- Rank the sample sums from smallest to largest. Ties are dealt by using average ranks.
- Let  $R_0$  be the rank of the observed sample sum with respect all the sample sums.

$$\alpha = 1 - \frac{R_0 - 0.5}{M_j}$$

b)  $M_j > 1,000$

- Draw a random sample of 1,000 sample sums from the permutation distribution.
- Add the observed sample sum to the list. There is a total of 1001 sample sums. Rank the sample sums from smallest to largest. Ties are dealt by using average ranks.
- Let  $R_0$  be the rank of the observed sample sum with respect all the sample sums.

$$\alpha = 1 - \frac{R_0 - 0.5}{1001}$$

#### *Proportion Measure*

$$Z_j = \frac{n_j a_{1j} - n_{1j} a_j}{\sqrt{\frac{n_{1j} n_{2j} a_j (n_j - a_j)}{n_j - 1}}}$$

#### *Rate Measure*

$$Z_j = \frac{n_{1j} - n_j q_j}{\sqrt{n_j q_j (1 - q_j)}}$$

### Ratio Measure

$$Z_j = \frac{\hat{R}_{1j} - \hat{R}_{2j}}{\sqrt{V(\hat{R}_{1j}) \left( \frac{1}{n_{1j}} + \frac{1}{n_{2j}} \right)}}$$

$$V(\hat{R}_{1j}) = \frac{\sum_k (U_{1jk} - \hat{R}_{1j} X_{1jk})^2}{\bar{X}_{1j}^2 (n_{1j} - 1)} = \frac{\sum_k U_{1jk}^2 - 2\hat{R}_{1j} \sum_k (U_{1jk} X_{1jk}) + \hat{R}_{1j}^2 \sum_k X_{1jk}^2}{\bar{X}_{1j}^2 (n_{1j} - 1)}$$

3. **Obtain a truncated Z value for each cell,  $Z_j^*$ .** To limit the amount of cancellation that takes place between cell results during aggregation, cells whose results suggest possible favoritism are left alone. Otherwise the cell statistic is set to zero. This means that positive equivalent Z values are set to 0, and negative values are left alone. Mathematically, this is written as

$$Z_j^* = \min(0, Z_j).$$

4. **Calculate the theoretical mean and variance of the truncated statistic under the null hypothesis of parity,  $E(Z_j^* | H_0)$  and  $\text{Var}(Z_j^* | H_0)$ .** In order to compensate for the truncation in step 3, an aggregated, weighted sum of the  $Z_j^*$  will need to be centered and scaled properly so that the final aggregate statistic follows a standard normal distribution.

- If  $W_j = 0$ , then no evidence of favoritism is contained in the cell. The formulae for calculating  $E(Z_j^* | H_0)$  and  $\text{Var}(Z_j^* | H_0)$  cannot be used. Set both equal to 0.
- If  $\min(n_{1j}, n_{2j}) > 6$  for a mean measure,  $\min\left\{a_{1j}\left(1 - \frac{a_{1j}}{n_{1j}}\right), a_{2j}\left(1 - \frac{a_{2j}}{n_{2j}}\right)\right\} > 9$  for a proportion measure,  $\min(n_{1j}, n_{2j}) > 15$  and  $n_j q_j (1 - q_j) > 9$  for a rate measure, or  $n_{1j}$  and  $n_{2j}$  are large for a ratio measure then

$$E(Z_j^* | H_0) = -\frac{1}{\sqrt{2\pi}}, \text{ and}$$

$$\text{Var}(Z_j^* | H_0) = \frac{1}{2} - \frac{1}{2\pi}.$$

- Otherwise, determine the total number of values for  $Z_j^*$ . Let  $z_{ji}$  and  $\theta_{ji}$ , denote the values of  $Z_j^*$  and the probabilities of observing each value, respectively.

$$E(Z_j^* | H_0) = \sum_i \theta_{ji} z_{ji}, \text{ and}$$

$$\text{Var}(Z_j^* | H_0) = \sum_i \theta_{ji} z_{ji}^2 - [E(Z_j^* | H_0)]^2.$$

The actual values of the  $z$ 's and  $\theta$ 's depends on the type of measure.

#### *Mean Measure*

$$N_j = \min(M_j, 1,000), \quad i = 1, K, N_j$$

$$z_{ji} = \min \left\{ 0, \Phi^{-1} \left( 1 - \frac{R_i - 0.5}{N_j} \right) \right\} \quad \text{where } R_i \text{ is the rank of sample sum } i$$

$$\theta_j = \frac{1}{N_j}$$

#### *Proportion Measure*

$$z_{ji} = \min \left\{ 0, \frac{n_j i - n_{1j} a_j}{\sqrt{\frac{n_{1j} n_{2j} a_j (n_j - a_j)}{n_j - 1}}} \right\}, \quad i = \max(0, a_j - n_{2j}), K, \min(a_j, n_{1j})$$

$$\theta_{ji} = \text{HG}(i)$$

#### *Rate Measure*

$$z_{ji} = \min \left\{ 0, \frac{i - n_j q_j}{\sqrt{n_j q_j (1 - q_j)}} \right\}, \quad i = 0, K, n_j$$

$$\theta_{ji} = \text{BN}(i)$$

#### *Ratio Measure*

The performance measure that is in this class is billing accuracy. The sample sizes for this measure are quite large, so there is no need for a small sample technique. If one does need a small sample technique, then a resampling method can be used.

5. Calculate the aggregate test statistic,  $Z^T$ .

$$Z^T = \frac{\sum_j W_j Z_j^* - \sum_j W_j E(Z_j^* | H_0)}{\sqrt{\sum_j W_j^2 \text{Var}(Z_j^* | H_0)}}$$

### Decision Process

Once  $Z^T$  has been calculated, it is compared to a critical value to determine if the ILEC is favoring its own customers over a CLEC's customers. The derivation of the critical value is found in Appendix C.

This critical value changes as the ILEC and CLEC transaction volume change. One way to make this transparent to the decision maker, is to report the difference between the test statistic and the critical value,  $diff = Z^T - c_B$ . If favoritism is concluded when  $Z^T < c_B$ , then the  $diff < 0$  indicates favoritism.

This make it very easy to determine favoritism: a positive  $diff$  suggests no favoritism, and a negative  $diff$  suggests favoritism. Appendix D provides an example of how this information can be reported for each month.



## **Appendix B. Trunk Blocking**

This Appendix provides an example of how the trunk blocking data can be processed to apply the Truncated Z Statistic. Trunk blocking is defined as the proportion of blocked calls a trunk group experiences in a time interval. It is a ratio of two numbers—blocked and attempted calls, both of which can vary over time and across trunk groups. Since the measure is a proportion where the numerator is a subset of the denominator, the truncated Z statistic, modified for proportions, can be applied here (see Appendix A).

As with other performance measures, data are first assigned to like-to-like cells, and the Z statistic is then computed within each cell. For trunk blocking, cells are defined by three variables: hour, day, and trunk group size or capacity. The next sections will describe the data and the data processing steps in greater detail.

The approach used in this example needs to be reviewed by subject matter expert to determine if it proper to use for trunk blocking.

### **Data Sources**

Two data files are processed for the trunk blocking measure. One is the Trunk Group Data File that contains the Trunk Group Serial Number (TGSN), Common Language Location Identifier (CLLI) , and other characteristics needed to categorize trunk groups and to identify them as BellSouth or CLEC.

The other file is the Blocking Data File (BDF), which contains the actual 24 hour blocking ratios for each weekday. There are 4 or 5 weeks in a monthly report cycle. The current system, however, allows the storage of daily blocking data by hour for a week only. Therefore, the data elements necessary to compute the Truncated Z must be extracted each week.

Two important data fields of interest on the Blocking Data File are the Blocking Ratio and Offered Load. The basic definition of Blocking Ratio is the proportion of all attempted calls that were blocked. For the simplest case of one way trunk groups, this is computed by dividing the number of blocked calls by the total call attempts, given that the data are valid. If they are not valid (e.g., actual usage exceeds capacity), blocking is estimated via the Neal Wilkinson algorithm.

Although the raw data--blocked calls (overflow) and peg counts (total call attempts)--are available, the calculation of the Blocking Ratio may be complicated for two-way trunk groups and trunk groups with invalid data. For this reason, we use the blocking ratios from the BDF instead of computing the ratios from the raw data. In order to reflect different call volumes processed through each trunk group, however, the blocking ratios need to be either weighted by call volume or converted to blocked and attempted calls before they are aggregated.

The measure of call traffic volume recommended for weighting is Offered Load. Offered Load is different from call counts in that it incorporates call duration as well. Since it is not just the number of calls but the total usage—number of calls multiplied by average call duration—that determines the occurrence of any blocking, this pseudo measure, Offered Load, appears to be the best indicator of call volume.

Cells or comparison classes are determined by three factors—hour, day, and trunk group capacity (number of trunks in service). The first two factors represent natural classes because trunk blocking changes over time. The third factor is based on our finding that high blocking tends to occur in small trunk groups. A pattern was found not only in the magnitude of blocking but also in its variability. Both the magnitude and variability of blocking decrease as trunk group capacity increases. Additional work is needed to establish the appropriate number of capacity levels and the proper location of boundaries.

## Data Processing

The data are processed using the five steps below:

1. Merge the two files by TGSN and select only trunk groups listed in both files.
2. Reset the blocking of all high use trunk groups to zero<sup>1</sup>.
3. Assign trunk group categories to CLEC and BellSouth: Categories 1, 3, 4, 5, 10, and 16 for CLEC and 9 for BellSouth<sup>2</sup>. The categories used here for comparison are:

Category	Administrator	Point A	Point B
1	BellSouth	BellSouth End Office	BellSouth Access Tandem
3	BellSouth	BellSouth End Office	CLEC Switch
4	BellSouth	BellSouth Local Tandem	CLEC Switch
5	BellSouth	BellSouth Access Tandem	CLEC Switch
9	BellSouth	BellSouth End Office	BellSouth End Office
10	BellSouth	BellSouth End Office	BellSouth Local Tandem
16	BellSouth	BellSouth Tandem	BellSouth Tandem

4. Recode the missing data. The Blocking Data File assigns all missing data (no valid measurement data) zero blocking. To differentiate true zero blocking from zeroes due to missing data, invalid records were identified and the ratios reset to missing. The blocking value was invalid if both the number of Loaded Days and the Offered Load were 0 for a given hourly period.
5. Form comparison classes based either on the data (i.e., quartiles) or on a predetermined set of values.

<sup>1</sup> The high use trunk groups cannot have any blocking. These are set up such that all overflow calls are automatically routed to other trunk groups instead of being physically blocked.

<sup>2</sup> More detailed information on all categories is described in a report 'Trunk Performance Report Generation' by Ernst & Young (March 1999).

## Calculation of the Proportion of Blocked Calls

Each cell is determined by day of the month, hour of the day, and trunk group capacity. To use the Truncated Z method, we generate summary information, to include the total number of blocked calls and the total number of attempted calls, for each cell.

For the details of each calculation step, the following notation is used. For a given hour of a day, let  $\bar{X}_{1ij}$  be the proportion of BellSouth blocked calls for trunk group i in cell j and  $\bar{X}_{2ij}$  be the corresponding proportion for CLEC. Then  $\bar{X}_{1ij} = X_{1ij} / n_{1ij}$  where  $X_{1ij}$  denotes the number of BellSouth blocked calls and  $n_{1ij}$  denotes the number of BellSouth total call attempts (indicated by Offered Load) for trunk group i in cell j. Likewise,  $\bar{X}_{2ij} = X_{2ij} / n_{2ij}$ . For the steps outlined below, only the CLEC notation is provided.

1. Compute the number of blocked calls for trunk group i:  $X_{2ij} = \bar{X}_{2ij} * n_{2ij}$
2. Compute total call attempts for all trunk groups in the cell:  $n_{2j} = \sum_i n_{2ij}$
3. Compute mean blocking proportion for cell j:  $\bar{X}_{2j} = \sum_i X_{2ij} / \sum_i n_{2ij}$
4. Compute the total number of BellSouth and CLEC blocked calls in cell j:  $t_j = \sum_i X_{1ij} + \sum_i X_{2ij}$
5. Apply the Truncated Z Statistic for Proportion measures presented in Appendix A.

## Appendix C

### Balancing the Type I and Type II Error Probabilities of the Truncated Z Test Statistic

This appendix describes a the methodology for balancing the error probabilities when the Truncated Z statistic, described in Appendix A, is used for performance measure parity testing. There are four key elements of the statistical testing process:

1. the null hypothesis,  $H_0$ , that parity exists between ILEC and CLEC services
2. the alternative hypothesis,  $H_a$ , that the ILEC is giving better service to its own customers
3. the Truncated Z test statistic,  $Z^T$ , and
4. a critical value,  $c$

The decision rule<sup>1</sup> is

- If  $Z^T < c$  then accept  $H_a$ .
- If  $Z^T \geq c$  then accept  $H_0$ .

There are two types of error possible when using such a decision rule:

**Type I Error:** Deciding favoritism exists when there is, in fact, no favoritism.

**Type II Error:** Deciding parity exists when there is, in fact, favoritism.

The probabilities of each type of each are:

**Type I Error:**  $\alpha = P(Z^T < c | H_0)$ .

**Type II Error:**  $\beta = P(Z^T \geq c | H_a)$ .

In what follows, we show how to find a balancing critical value,  $c_B$ , so that  $\alpha = \beta$ .

#### General Methodology

The general form of the test statistic that is being used is

---

<sup>1</sup> This decision rule assumes that a negative test statistic indicates poor service for the CLEC customer. If the opposite is true, then reverse the decision rule.

$$z_0 = \frac{\hat{T} - E(\hat{T} | H_0)}{SE(\hat{T} | H_0)}, \quad (C.1)$$

where

$\hat{T}$  is an estimator that is (approximately) normally distributed,

$E(\hat{T} | H_0)$  is the expected value (mean) of  $\hat{T}$  under the null hypothesis, and

$SE(\hat{T} | H_0)$  is the standard error of  $\hat{T}$  under the null hypothesis.

Thus, under the null hypothesis,  $z_0$  follows a standard normal distribution. However, this is not true under the alternative hypothesis. In this case,

$$z_a = \frac{\hat{T} - E(\hat{T} | H_a)}{SE(\hat{T} | H_a)}$$

has a standard normal distribution. Here

$E(\hat{T} | H_a)$  is the expected value (mean) of  $\hat{T}$  under the alternative hypothesis, and

$SE(\hat{T} | H_a)$  is the standard error of  $\hat{T}$  under the alternative hypothesis.

Notice that

$$\begin{aligned} \beta &= P(z_0 > c | H_a) \\ &= P\left(z_a > \frac{cSE(\hat{T} | H_0) + E(\hat{T} | H_0) - E(\hat{T} | H_a)}{SE(\hat{T} | H_a)}\right) \end{aligned} \quad (C.2)$$

and recall that for a standard normal random variable  $z$  and a constant  $b$ ,  $P(z < b) = P(z > -b)$ . Thus,

$$\alpha = P(z_0 < c) = P(z_0 > -c) \quad (C.3)$$

Since we want  $\alpha = \beta$ , the right hand sides of (C.2) and (C.3) represent the same area under the standard normal density. Therefore, it must be the case that

$$-c = \frac{cSE(\hat{T} | H_0) + E(\hat{T} | H_0) - E(\hat{T} | H_a)}{SE(\hat{T} | H_a)}.$$

Solving this for  $c$  gives the general formula for a balancing critical value:

$$c_B = \frac{E(\hat{T} | H_a) - E(\hat{T} | H_0)}{SE(\hat{T} | H_a) + SE(\hat{T} | H_0)} \quad (C.4)$$

### The Balancing Critical Value of the Truncated Z

In Appendix A, the Truncated Z statistic is defined as

$$Z^T = \frac{\sum_j W_j Z_j^* - \sum_j W_j E(Z_j^* | H_0)}{\sqrt{\sum_j W_j^2 \text{Var}(Z_j^* | H_0)}}$$

In terms of equation (C.1) we have

$$\begin{aligned} \hat{T} &= \sum_j W_j Z_j^* \\ E(\hat{T} | H_0) &= \sum_j W_j E(Z_j^* | H_0) \\ SE(\hat{T} | H_0) &= \sqrt{\sum_j W_j^2 \text{Var}(Z_j^* | H_0)} \end{aligned}$$

To compute the balancing critical value (C.4), we also need  $E(\hat{T} | H_a)$  and  $SE(\hat{T} | H_a)$ . These values are determined by

$$\begin{aligned} E(\hat{T} | H_a) &= \sum_j W_j E(Z_j^* | H_a), \text{ and} \\ SE(\hat{T} | H_a) &= \sqrt{\sum_j W_j^2 \text{var}(Z_j^* | H_a)}. \end{aligned}$$

In which case equation (C.4) gives

$$c_B = \frac{\sum_j W_j E(Z_j^* | H_a) - \sum_j W_j E(Z_j^* | H_0)}{\sqrt{\sum_j W_j^2 \text{var}(Z_j^* | H_a) + \sum_j W_j^2 \text{var}(Z_j^* | H_0)}}. \quad (C.5)$$

Thus, we need to determine how to calculate  $E(Z_j^* | H_0)$ ,  $\text{Var}(Z_j^* | H_0)$ ,  $E(Z_j^* | H_a)$ , and  $\text{Var}(Z_j^* | H_a)$ .

If  $Z_j$  has a normal distribution with mean  $\mu$  and standard error  $\sigma$ , then the mean of the distribution truncated at 0 is

$$M(\mu, \sigma) = \int_{-\infty}^0 \frac{x}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx,$$

and the variance is

$$V(\mu, \sigma) = \int_{-\infty}^0 \frac{x^2}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx - M(\mu, \sigma)^2$$

It can be shown that

$$M(\mu, \sigma) = \mu \Phi\left(\frac{-\mu}{\sigma}\right) - \sigma \phi\left(\frac{-\mu}{\sigma}\right)$$

and

$$V(\mu, \sigma) = (\mu^2 + \sigma^2) \Phi\left(\frac{-\mu}{\sigma}\right) - \mu \sigma \phi\left(\frac{-\mu}{\sigma}\right) - M(\mu, \sigma)^2$$

where  $\Phi(\cdot)$  is the cumulative standard normal distribution function, and  $\phi(\cdot)$  is the standard normal density function.

The cell test statistic,  $Z_j$ , is constructed so that it has mean 0 and standard deviation 1 under the null hypothesis. Thus,

$$E(Z_j^* | H_0) = M(0, 1) = -\frac{1}{\sqrt{2\pi}}, \text{ and}$$

$$\text{var}(Z_j^* | H_0) = V(0, 1) = \frac{1}{2} - \frac{1}{2\pi}.$$

The mean and standard error of  $Z_j$  under the alternative hypothesis depends on the type of measure and the form of the alternative. These are discussed below. For now, denote the mean and standard error of  $Z_j$  under the alternative by  $m_j$  and  $se_j$  respectively. Thus,

$$E(Z_j^* | H_a) = M(m_j, se_j), \text{ and}$$

$$SE(Z_j^* | H_a) = V(m_j, se_j).$$

Using the above notation, and equation (C.5), we get the formula for the balancing critical of  $Z^*$ .

$$c_R = \frac{\sum_j W_j M(m_j, se_j) - \sum_j W_j \frac{-1}{\sqrt{2\pi}}}{\sqrt{\sum_j W_j^2 V(m_j, se_j) + \sum_j W_j^2 \left( \frac{1}{2} - \frac{1}{2\pi} \right)}}. \quad (C.6)$$

This formula assumes that  $Z_j$  is approximately normally distributed within cell  $j$ . When the cell sample sizes,  $n_{1j}$  and  $n_{2j}$ , are small this may not be true. It is possible to determine the cell mean and variance under the null hypothesis when the cell sample sizes are small. It is much more difficult to determine these values under the alternative hypothesis. Since the cell weight,  $W_j$ , will also be small (see Appendix A) for a cell with small volume, the cell mean and variance will not contribute much to the weighted sum. Therefore, formula (C.6) provides a reasonable approximation to the balancing critical value.

### Alternative Hypotheses

#### Mean Measure

For mean measures, one is concerned with two parameters in each cell, namely, the mean and variance. A possible lack of parity may be due to a difference in cell means, and/or a difference in cell variances. One possible set of hypotheses that capture this notion, and take into account the assumption that transactions are identically distributed within cells is:

$$H_0: \mu_{1j} = \mu_{2j}, \sigma_{1j}^2 = \sigma_{2j}^2$$

$$H_a: \mu_{2j} = \mu_{1j} + \delta_j \cdot \sigma_{1j}, \sigma_{2j}^2 = \lambda_j \cdot \sigma_{1j}^2 \quad \delta_j > 0, \lambda_j \geq 1 \text{ and } j = 1, \dots, L.$$

Under this form of alternative hypothesis, the cell test statistic  $Z_j$  has mean and standard error given by

$$m_j = \frac{-\delta_j}{\sqrt{\frac{1}{n_{1j}} + \frac{1}{n_{2j}}}}, \text{ and}$$

$$se_j = \sqrt{\frac{\lambda_j n_{1j} + n_{2j}}{n_{1j} + n_{2j}}}$$

#### Proportion Measure

For a proportion measure there is only one parameter of interest in each cell, the proportion of transactions possessing an attribute of interest. A possible lack of parity may be due to a difference in cell proportions. A set of hypotheses that take into account



the assumption that transaction are identically distributed within cells while allowing for an analytically tractable solution is:

$$H_0: \frac{p_{2j}(1-p_{1j})}{(1-p_{2j})p_{1j}} = 1$$

$$H_a: \frac{p_{2j}(1-p_{1j})}{(1-p_{2j})p_{1j}} = \psi_j \quad \psi_j > 1 \text{ and } j = 1, \dots, L.$$

These hypotheses are based on the “odds ratio.” If the transaction attribute of interest is a missed trouble repair, then an interpretation of the alternative hypothesis is that a CLEC trouble is  $\psi_j$  times more likely to be missed than an ILEC trouble.

Under this form of alternative hypothesis, the within cell asymptotic mean and variance of  $a_{ij}$  are given by<sup>2</sup>

$$E(a_{ij}) = n_j \pi_j^{(1)}$$

$$\text{var}(a_{ij}) = \frac{n_j}{\frac{1}{\pi_j^{(1)}} + \frac{1}{\pi_j^{(2)}} + \frac{1}{\pi_j^{(3)}} + \frac{1}{\pi_j^{(4)}}} \quad (C.7)$$

where

$$\pi_j^{(1)} = f_j^{(1)} \left( n_j^2 + f_j^{(2)} + f_j^{(3)} - f_j^{(4)} \right)$$

$$\pi_j^{(2)} = f_j^{(1)} \left( -n_j^2 - f_j^{(2)} + f_j^{(3)} + f_j^{(4)} \right)$$

$$\pi_j^{(3)} = f_j^{(1)} \left( -n_j^2 + f_j^{(2)} - f_j^{(3)} + f_j^{(4)} \right)$$

$$\pi_j^{(4)} = f_j^{(1)} \left( n_j^2 \left( \frac{2}{\psi_j} - 1 \right) - f_j^{(2)} - f_j^{(3)} - f_j^{(4)} \right)$$

$$f_j^{(1)} = \frac{1}{2n_j^2 \left( \frac{1}{\psi_j} - 1 \right)}$$

$$f_j^{(2)} = n_j n_{1j} \left( \frac{1}{\psi_j} - 1 \right)$$

$$f_j^{(3)} = n_j a_j \left( \frac{1}{\psi_j} - 1 \right)$$

$$f_j^{(4)} = \sqrt{n_j^2 \left[ 4n_{1j} (n_j - a_j) \left( \frac{1}{\psi_j} - 1 \right) + \left( n_j + (a_j - n_{1j}) \left( \frac{1}{\psi_j} - 1 \right) \right)^2 \right]}$$

<sup>2</sup> Stevens, W. L. (1951) Mean and Variance of an entry in a Contingency Table. *Biometrika*, 38, 468-470.

Recall that the cell test statistic is given by

$$Z_j = \frac{n_j a_{1j} - n_{1j} a_j}{\sqrt{\frac{n_{1j} n_{2j} a_j (n_j - a_j)}{n_j - 1}}}.$$

Using the equations in (C.7), we see that  $Z_j$  has mean and standard error given by

$$m_j = \frac{n_j^2 \pi_j^{(1)} - n_{1j} a_j}{\sqrt{\frac{n_{1j} n_{2j} a_j (n_j - a_j)}{n_j - 1}}}, \text{ and}$$

$$se_j = \sqrt{\frac{n_j^3 (n_j - 1)}{n_{1j} n_{2j} a_j (n_j - a_j) \left( \pi_j^{(1)} + \pi_j^{(2)} + \pi_j^{(3)} + \pi_j^{(4)} \right)}}.$$

#### *Rate Measure*

A rate measure also has only one parameter of interest in each cell, the rate at which a phenomenon is observed relative to a base unit, e.g. the number of troubles per available line. A possible lack of parity may be due to a difference in cell rates. A set of hypotheses that take into account the assumption that transaction are identically distributed within cells is:

$$H_0: r_{1j} = r_{2j}$$

$$H_a: r_{2j} = \epsilon_j r_{1j} \quad \epsilon_j > 1 \text{ and } j = 1, \dots, L.$$

Given the total number of ILEC and CLEC transactions in a cell,  $n_j$ , and the number of base elements,  $b_{1j}$  and  $b_{2j}$ , the number of ILEC transaction,  $n_{1j}$ , has a binomial distribution from  $n_j$  trials and a probability of

$$q_j^* = \frac{r_{1j} b_{1j}}{r_{1j} b_{1j} + r_{2j} b_{2j}}.$$

Therefore, the mean and variance of  $n_{1j}$ , are given by

$$\begin{aligned} E(n_{1j}) &= n_j q_j^* \\ \text{var}(n_{1j}) &= n_j q_j^* (1 - q_j^*) \end{aligned} \tag{C.8}$$

Under the null hypothesis

$$q_j^* = q_j = \frac{b_{1j}}{b_j},$$

but under the alternative hypothesis

$$q_j^* = q_j^a = \frac{b_{1j}}{b_{1j} + \epsilon_j b_{2j}}. \quad (C.9)$$

Recall that the cell test statistic is given by

$$Z_j = \frac{n_{1j} - n_j q_j}{\sqrt{n_j q_j (1 - q_j)}}.$$

Using (C.8) and (C.9), we see that  $Z_j$  has mean and standard error given by

$$m_j = \frac{n_j (q_j^a - q_j)}{\sqrt{n_j q_j (1 - q_j)}} = (1 - \epsilon_j) \sqrt{\frac{n_j b_{1j} b_{2j}}{b_{1j} + \epsilon_j b_{2j}}}, \text{ and}$$

$$se_j = \sqrt{\frac{q_j^a (1 - q_j^a)}{q_j (1 - q_j)}} = \sqrt{\epsilon_j} \frac{b_j}{b_{1j} + \epsilon_j b_{2j}}.$$

### *Ratio Measure*

As with mean measures, one is concerned with two parameters in each cell, the mean and variance, when testing for parity of ratio measures. As long as sample sizes are large, as in the case of billing accuracy, the same method for finding  $m_j$  and  $se_j$  that is used for mean measures can be used for ratio measures.

### **Determining the Parameters of the Alternative Hypothesis**

In this appendix we have indexed the alternative hypothesis of mean measures by two sets of parameters,  $\lambda_j$  and  $\delta_j$ . Proportion and rate measures have been indexed by one set of parameters each,  $\psi_j$  and  $\epsilon_j$  respectively. A major difficulty with this approach is that more than one alternative will be of interest; for example we may consider one alternative in which all the  $\delta_j$  are set to a common non-zero value, and another set of alternatives in each of which just one  $\delta_i$  is non-zero, while all the rest are zero. There are very many other possibilities. Each possibility leads to a single value for the balancing critical value; and each possible critical value corresponds to many sets of alternative hypotheses, for each of which it constitutes the correct balancing value.

The formulas we have presented can be used to evaluate the impact of different choices of the overall critical value. For each putative choice, we can evaluate the set of alternatives for which this is the correct balancing value. While statistical science can be used to evaluate the impact of different choices of these parameters, there is not much that an appeal to statistical principles can offer in directing specific choices. Specific choices are best left to telephony experts. Still, it is possible to comment on some aspects of these choices:

- Parameter Choices for  $\lambda_j$ . The set of parameters  $\lambda_j$  index alternatives to the null hypothesis that arise because there might be greater unpredictability or variability in the delivery of service to a CLEC customer over that which would be achieved for an otherwise comparable ILEC customer. While concerns about differences in the variability of service are important, it turns out that the truncated Z testing which is being recommended here is relatively insensitive to all but very large values of the  $\lambda_j$ . Put another way, reasonable differences in the values chosen here could make very little difference in the balancing points chosen.
- Parameter Choices for  $\delta_j$ . The set of parameters  $\delta_j$  are much more important in the choice of the balancing point than was true for the  $\lambda_j$ . The reason for this is that they directly index differences in average service. The truncated Z test is very sensitive to any such differences; hence, even small disagreements among experts in the choice of the  $\delta_j$  could be very important. Sample size matters here too. For example, setting all the  $\delta_j$  to a single value –  $\delta_j = \delta$  – might be fine for tests across individual CLECs where currently in Louisiana the CLEC customer bases are not too different. Using the same value of  $\delta$  for the overall state testing does not seem sensible. At the state level we are aggregating over CLECs, so using the same  $\delta$  as for an individual CLEC would be saying that a "meaningful" degree of disparity is one where the violation is the same ( $\delta$ ) for each CLEC. But the detection of disparity for any component CLEC is important, so the relevant "overall"  $\delta$  should be smaller.
- Parameter Choices for  $\psi_j$  or  $\epsilon_j$ . The set of parameters  $\psi_j$  or  $\epsilon_j$  are also important in the choice of the balancing point for tests of their respective measures. The reason for this is that they directly index increases in the proportion or rate of service performance. The truncated Z test is sensitive to such increases; but not as sensitive as the case of  $\delta$  for mean measures. Sample size matters here too. As with mean measures, using the same value of  $\psi$  or  $\epsilon$  for the overall state testing does not seem sensible.

The three parameters are related however. If a decision is made on the value of  $\delta$ , it is possible to determine equivalent values of  $\psi$  and  $\epsilon$ . The following equations, in conjunction with the definitions of  $\psi$  and  $\epsilon$ , show the relationship with delta.

$$\delta = 2 \cdot \arcsin(\sqrt{\hat{p}_2}) - 2 \cdot \arcsin(\sqrt{\hat{p}_1})$$

$$\delta = 2\sqrt{\hat{t}_2} - 2\sqrt{\hat{t}_1}$$

The bottom line here is that beyond a few general considerations, like those given above, a principled approach to the choice of the alternative hypotheses to guard against must come from elsewhere.

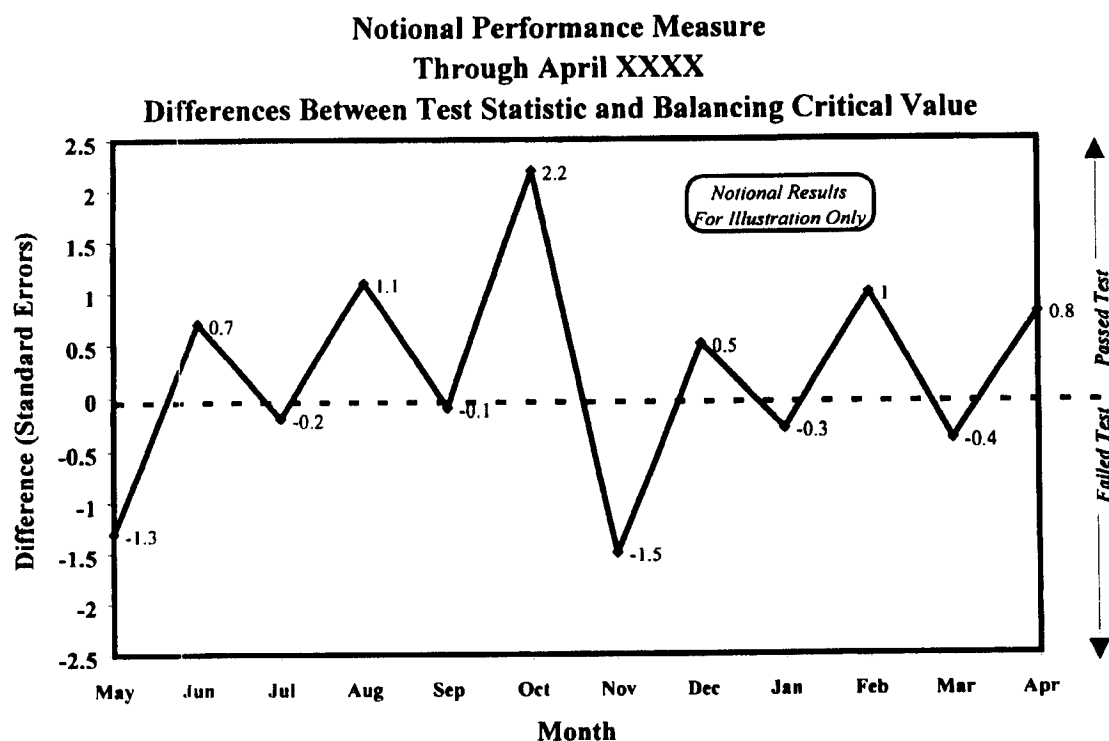
## **Appendix D: Examples of Statistical Reports**

The general structure for reporting statistical results in a production environment will be the same for the different measures and we suggest that it consist of at least three components. For each measure present, (1) the monthly test statistics over a period of time, (2) the results for the current month, with summary statistics, test statistics, and descriptive graphs, and (3) a summary of any adjustments to the data made in the process of running the tests, including a description of how many records were excluded from analysis and the reason for the exclusion (i.e., excluded due to business rules, or due to statistical/methodological rules pertaining to the measure). The last component is important to assure that the reported results can be audited.

Selected components of the reporting structure are illustrated in the samples that follow. An outline of the report is shown below. Monthly results will be presented for each level of aggregation required.

- I. Test Statistics Over Time
- II. Monthly Results
  - A. Summary Statistics
  - B. Test Statistics
  - C. Descriptive Graphs (Frequency Distributions, etc.)
- III. Adjustments to Data
  - A. Records Excluded Due to Business Rules
  - B. Records Excluded Due to Statistical Rules

**Test Statistic Over Time.** The first component of the reporting structure is an illustration of the trend of the particular performance measure over time together with a tabular summary of results for the current month. We will show at a glance whether the tests consistently return non-statistically significant results; consistently indicate disparity (be that in favor of BellSouth or in favor of the CLECs); or vary month by month in their results. An example of this component follows.

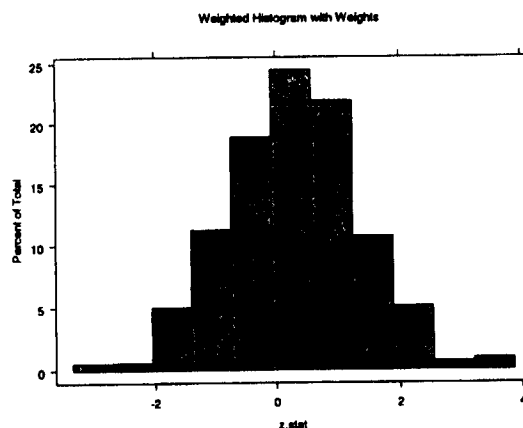
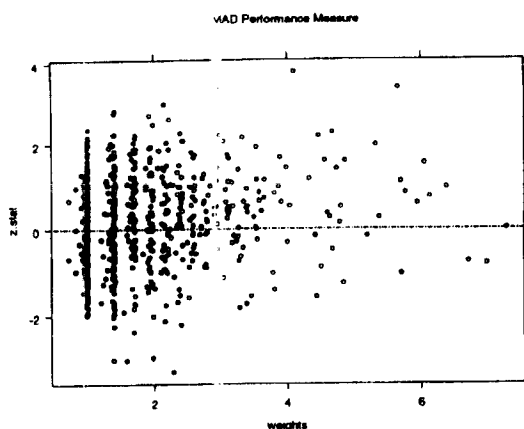


Result for Current Month	
Test Statistic	-0.410
Balancing Critical Value	-1.210
Difference	0.800

**Monthly Results.** The most important component of the reporting structure is the part which presents results of the monthly statistical tests on the given performance measure. The essential aspects included in this component are the summary statistics; the test statistics and results; and descriptive graphs of the results.

It is important to present basic summary statistics to complete the comparison between BellSouth and the CLECs. At a minimum, these statistics will include the means, standard deviations, and population sizes. In addition to basic descriptive statistics, we also present the test statistic results. Examples of ways we have presented these statistics in the past can be found in BellSouth's February 25, 1999 filing before the Louisiana Public Service Commission.

Finally, the results will be presented in graphical format. Below is an example of how to graphically present the data behind the Truncated Z statistic. One graph shows a plot of cell Z score versus cell weights. The other is a histogram of the weighted cell Z scores.



**Adjustments to Data.** The third important component of the reporting structure is information on any adjustments performed on the data. This information is essential in order that the results may be verified and audited. The most prevalent examples of such modifications would be removal of observations and weighting of the data.

Records can be removed from analysis for both business reasons (these will likely be taken into account in the PMAP system) and for statistical reasons. All of the performance measures exclude certain records based on business rules underlying each measure's particular definitions and methodologies. The number of records excluded for each rule will be summarized. In addition, some of the measures will have observations excluded for statistical reasons, particularly in the case of "mean measures" (OCI and MAD); these exclusions will be summarized as well. The tables below show examples of the current method for summarizing this information:

# April XXXX

## Perormance Measure Filtering Information

This table displays information about the size of the database files and the cases that were removed from the analysis.

<b>Unfiltered Total</b>	<b>28,691</b>
<b>Records Removed for Business Reasons</b> <i>(e.g. not N, T, C, or P orders, not resale and not UNE)</i>	<b>7,242</b>
<b>Total Reported on Web Report</b>	<b>21,449</b>
<b>Additional Records Removed for Business Reasons</b>	<b>876</b>
Missing Appointment code is 'S'	844
General Class Service = 'O'	0
UNE Cases	102
<b>Records Removed for Statistical Reasons</b>	
Extremc Values Removed	9
<b>No Matching Classification Removals</b>	<b>47</b>
<b>FILTERED TOTAL</b>	<b>20,517</b>

1999

<b>Unfiltered Total</b>	<b>453,107</b>
<b>Records Removed for Business Reasons</b> <i>(e.g. not N, T, C, or P orders, not retail)</i>	<b>78,613</b>
<b>Total Reported on Web Report</b>	<b>374,494</b>
<b>Additional Records Removed for Business Reasons</b>	<b>7,429</b>
Missing Appointment code is 'S'	7,172
General Class Service = 'O'	279
<b>Records Removed for Statistical Reasons</b>	
Extremc Values Removed	652
<b>No Matching Classification Removals</b>	<b>21,974</b>
<b>FILTERED TOTAL</b>	<b>344,439</b>



## Appendix E. Trimming Outliers for Mean Measures

The arithmetic average is extremely sensitive to outliers; a single large value, possibly an erroneous value, can significantly distort the mean value. And by inflating the error variance, this also affects conclusions in the test of hypotheses. Extreme data values may be correct, but since they are rare measurements, they may be considered to be statistical outliers. Or they may be values that should not be in the analysis data set because of errors in the measurement or in selecting the data.

At this time, only two mean measures have been analyzed: Order Completion Interval and Maintenance Average Duration. Maintenance Average Duration data are truncated at 240 hours and therefore this measure was not trimmed further. For Order Completion Interval, the underlying distribution of the observations is clearly not normal, but rather skewed with a very long upper-tail.

A useful technique, coming from the field of robust statistical analysis, is to trim a very small proportion from the tails of the distribution before calculating the means. The resulting mean is referred to as a trimmed mean. Trimming is beneficial in that it speeds the convergence of the distribution of the means to a normal distribution. Only extreme values are trimmed, and in many cases the data being trimmed are, in fact, data that might not be used in the analysis on other grounds.

In the first analysis of the verified Order Completion Interval-Provisioning measure, after removing data that were clearly in error or were not applicable, we looked at the cases that represented the largest 0.01% of the BST distribution. In the August data, this corresponded to orders with completion intervals greater than 99 days. All of these were BellSouth orders. In examining the largest 11 individual examples that would be removed from analysis, we found that only 1 of the 11 cases was a valid case where the completion interval was unusually large. The other 10 cases were examples of cases that should not have been included in the analysis. This indicates that at least in preliminary analysis, it is both beneficial to examine the extreme outliers and reasonable to remove them.

A very slight trimming is needed in order to put the central limit theorem argument on firm ground. But finding a robust rule that can be used in a production setting is difficult. Also, any trimming rule should be fully explained and any observations that are trimmed from the data must be fully documented.

When it is determined that a measure should be trimmed, a trimming rule that is easy to implement in a production setting is:

**Trim the ILEC observations to the largest CLEC value from all CLEC observations in the month under consideration.**

That is, no CLEC values are removed; all ILEC observations greater than the largest CLEC observation are trimmed.

While this method is simple, it does allow for extreme CLEC observations to be part of the analysis. For instance, suppose that the amount of time to complete an order was less than 40 days for all CLEC orders except one. Let's say that this extreme order took 100 days to complete. The trimming rule says that all ILEC orders above 100 days should be trimmed, but a closer look at the data might suggest trimming at 40 days instead.

Since we are operating in a production mode system, it is not possible to explore the data before the trimming takes place. Other automatic trimming rules present other problems, so our solution is to use the simple trimming rule above, and have the system automatically produce a trimming report that can be examined at a later point in time.

The trimming report should include:

- The value of the trim point.
- Summary statistics and graphics of the ILEC observations that were trimmed.
- A listing of the trimmed ILEC transaction for a random sample of 10 trimmed transactions. This listing should not disclose sensitive information.
- A listing of the 10 most extreme CLEC transactions. This listing should not disclose sensitive information.
- The number of ILEC and CLEC observations above some fixed point, so that changes in the upper tail can be better tracked over time.

The trimming report should be part of the overall report discussed in Appendix D. Examples of tables contained within the trimming report are shown below.

**April XXXX**  
**Performance Measure Extreme Values**

<b>CLEC</b>		<b>ILEC</b>	
Cutoff	26	Cutoff	26
# of Records	20,573	# of Records	367,065
10 Largest		Extreme Values	652
Minimum	19	Minimum	27
Median	23	Median	32
Maximum	26	Maximum	283
<b>Subtotal</b>	<b>20,573</b>	<b>Subtotal</b>	<b>366,413</b>

**April XXXX**  
**Performance Measure Weighting Report**

<b>CLEC</b>		<b>ILEC</b>	
# of Records	20,573	# of Records	366,413
No Matching BST		No Matching CLEC	
Classification (1)	47	Classification (2)	21,974
<b>Subtotal</b>	<b>20,526</b>	<b>Subtotal</b>	<b>344,439</b>

**April XXXX**  
**Perormance Measure Filtering Information**

This table displays information about the size of the database files and the cases that were removed from the analysis.

		1999		
<b>Unfiltered Total</b>	<b>28,691</b>		<b>Unfiltered Total</b>	<b>453,107</b>
<b>Records Removed for Business Reasons</b> (e.g. not N, T, C, or P orders, not resale and not UNE)	<b>7,242</b>		<b>Records Removed for Business Reasons</b> (e.g. not N, T, C, or P orders, not retail)	<b>78,613</b>
<b>Total Reported on Web Report</b>	<b>21,449</b>		<b>Total Reported on Web Report</b>	<b>374,494</b>
<b>Additional Records Removed for Business Reasons</b>			<b>Additional Records Removed for Business Reasons</b>	
Missing Appointment code is 'S'	876		Missing Appointment code is 'S'	7,429
General Class Service = 'O'	844		General Class Service = 'O'	7,172
UNE Cases	0		General Class Service = 'O'	279
	102			
<b>Records Removed for Statistical Reasons</b>			<b>Records Removed for Statistical Reasons</b>	
Extreme Values Removed	0		Extreme Values Removed	652
<b>No Matching Classification Removals</b>	<b>47</b>		<b>No Matching Classification Removals</b>	<b>21,974</b>
<b>FILTERED TOTAL</b>	<b>20,526</b>		<b>FILTERED TOTAL</b>	<b>344,439</b>

**CLEC Extreme Values**

Wire Center	Time	Dispatch	Residence	Circuits	Order Type	Order Interval
NWORLAMA	1	1	3	1	N	61
OPLSLATL	1	2	1	1	C	53
NWORLAMA	2	1	3	1	N	44
NWORLAMA	1	1	3	1	N	39
BTRGLAWN	1	1	2	1	C	38
LKCHLADT	1	1	1	1	T	37
NWORLAMA	1	1	3	1	N	32
NWORLAMA	2	1	3	1	N	32
SHPTLACL	1	1	2	1	N	28

**Frequency of Extreme Values Removed from BST file (Top 10)**

Wire Center	Time	Dispatch	Residence	Circuits	Order Type	Frequency
NWORLAMA	1	1	3	1	N	55
NWORLAMA	2	1	3	1	N	25
BTRGLASB	2	1	3	1	C	23
NWORLAMC	2	1	3	1	C	23
NWORLAMC	1	1	3	1	C	22
NWORLAMA	2	1	3	1	C	18
NWORLAMA	1	1	3	1	C	17
BTRGLASB	1	1	3	1	C	16
LFYTLAMA	1	1	3	1	C	15
NWORLAMA	2	2	3	1	C	14

**Corrections**

LPSC “Statistical Techniques for the Analysis and Comparison of Performance Measure Data”,

Appendix A, page A-5

$$T_j = t_j + \frac{g}{6} \left( \frac{n_{1j} + 2n_{2j}}{\sqrt{n_{1j} n_{2j} (n_{1j} + n_{2j})}} \right) \left( t_j^2 + \frac{n_{2j} - n_{1j}}{n_{1j} + 2n_{2j}} \right)$$

Appendix C, page C-8, rate measures section for balancing critical value.

$$m_j = \frac{n_j (q_j^a - q_j)}{\sqrt{n_j q_j (1 - q_j)}} = (1 - \epsilon_j) \frac{\sqrt{n_j b_{1j} b_{2j}}}{b_{1j} + \epsilon_j b_{2j}}$$